Computer Aided Detection of Masses in Digital Mammograms

Een wetenschappelijke proeve op het
gebied van de Medische Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
dinsdag 11 januari 2000
des namiddags om 1.30 uur precies

door

Guido Maria te Brake
geboren 9 augustus 1969 te Utrecht

Promotor:	Prof. dr. C.C.A.M. Gielen
Copromotor:	Dr. ir. N. Karssemeijer
Manuscriptcommissie:	Prof. dr. A.L.M. Verbeek
	Dr. H.J. Kappen
	Prof. dr. J.N. Kok (UL)
	Prof. dr. J.H.J. Ruijs
	Prof. dr. ir. M.A. Viergever (UU)

ISBN 90-9013318-6

This research was made possible by Grant KUN 96-1343 from the Dutch Cancer Society

Grants for publication of this thesis by
- R2 Technology, Inc
- General Electric Medical Systems
- Siemens Nederland N.V.
- Kodak Nederland
- Stichting Radiologisch Onderzoek Nijmegen Rad-ON

are gratefully acknowledged.

Contents

Chapter 1

Introduction

Breast cancer is the most common cancer among women in the Western world. In the Netherlands, 9000 women are diagnosed with breast cancer annually, and 3500 women die from the disease. Men contribute less than 1% of annual incidence. The incidence of breast cancer in the Netherlands is among the highest in the world, just behind the United States. Approximately 1 out of each 10 women will develop breast cancer during her life.

The anatomy of the breast is very complex [62]. Each breast contains between 15 to 20 lobes that are connected to the nipple through a complex structure of converging ducts. Each lobule consists of 10 to 100 terminal duct lobular units (TDLU), the areas where breast cancer originates. When the tumor has not gone through the basal membrane but is completely contained in the lobule or the ducts the cancer is called *in situ*. When the cancer has broken trough the basal membrane it is called invasive, and chances on metastases increase sharply. The success of treatment of breast cancer depends largely on the stage at the time of detection. Two features determine the stage of a tumor: its size and whether metastases have been found in lymph nodes or distant areas. If the size of the tumor is smaller than 2 cm, preferable even under 1 cm, and if no lymph-nodes or distant areas are metastatic, chances of successful treatment are high. Therefore, in many countries breast cancer screening programs using mammography have been started to detect cancers as early as possible. Screening for breast cancer is a complex task, due to the large fraction of normal cases: only approximately 5 out of 1000 women have breast cancer. To help radiologists in their task to detect signs of cancer between large numbers of normal mammograms, a number of research groups are developing software for computer aided diagnosis (CAD). It is hoped that CAD can help to decrease the number of errors, both false negatives (malignant cases that were not recalled) and false positives (cases that are recalled unnecessarily). The topic of this thesis is automated detection of masses and architectural distortions in mammograms. This introduction will describe the possible uses of CAD in breast cancer screening programs, and will provide a context for the rest of this thesis.

1.1 Imaging modalities

Mammography, making images of the breast using X-rays, is the most widely used modality to detect and characterize breast cancer. Mammography has high sensitivity and specificity, even small tumors and microcalcifications can be detected on mammograms. The breast is

compressed between two plexiglas plates to flatten and immobilize it. Part of the X-rays is absorbed by the breast, part goes through the breast and hits a screen. This screen emits light when hit, blackening a film that is located just in front of it. Some X-rays exit the breast under a different angle, causing a blurring effect called scatter. Scatter can be reduced by positioning a grid in front of the screen that will absorb the scattered X-rays, but when a grid is used a higher dose is required. The film/screen systems will be replaced in the next years by digital detectors that have a higher dynamic range, and may require a lower X-ray dose.

Mammograms show a projection of the breast that can be made from different angles. The two most common projections are medio-lateral oblique and cranio-caudal, shown in Figure 1.1. The advantage of the medio-lateral oblique projection is that almost the whole breast is visible, often including lymph nodes. Part of the pectoral muscle will be shown in upper part of the image, which is superimposed over a portion of the breast. The cranio-caudal view is taken from above, resulting in an image that sometimes does not show the area close to the chest wall.

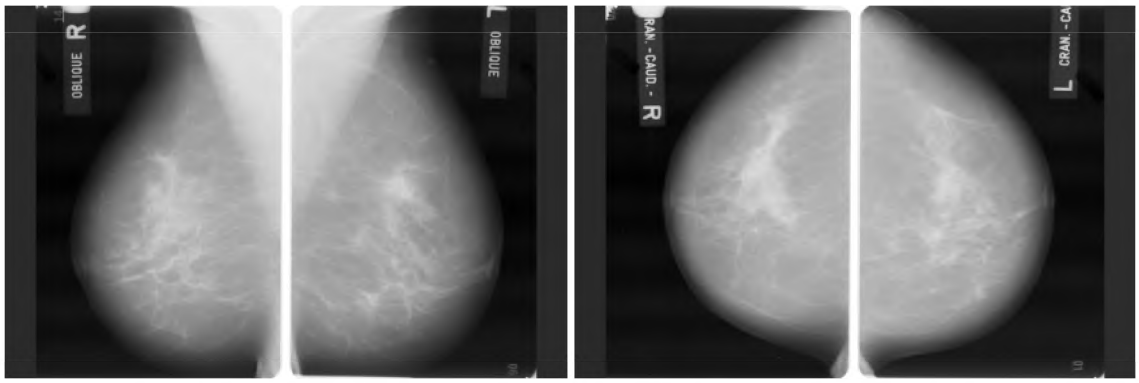

Figure 1.1: A mammogram, two oblique and two cranio-caudal films

Mammography does have a number of drawbacks. It is an invasive technique because the women is exposed to a (low) radiation dose. Also, the compression of the breast can be a painful experience.

Approximately 10% of all malignant abnormalities is radiographically occult and therefore not visible on mammograms. If palpable lesions are found that cannot be made visible using mammography, ultrasound or magnetic resonance imaging (MRI) can be used to further examine the breast. Ultrasound has rather low sensitivity and specificity and is not useful for screening, but has proven to be a useful modality for discriminating solid and cystic lesions [30]. Contrast-enhanced magnetic resonance imaging has shown to be useful for discriminating benign and malignant lesions, and due to its high sensitivity it is able to show a large number of lesion that are radiographically occult [6, 7]. MRI is not a useful modality for screening, because of its low specificity and high cost. Mammography is the only image modality suited for screening programs because of its high performance and low costs.

1.2 Mammographic signs of cancer

A number of signs in mammograms can indicate the presence of a malignant process. When the cancer is still *in situ*, the only sign that may appear on mammograms are microcalcifications, but several benign processes can cause microcalcifications as well. The shape and topology of these calcifications indicate whether they are the result of a benign or a malignant process. In Figure 1.2 an example of a microcalcification cluster is shown. Microcalcifications are small (between 0.1 mm and 2.0 mm) and can occur alone or in clusters. Classification of microcalcifications is important, because recalling all microcalcification clusters will result in many false positives since 80% of all clusters are due to benign processes.

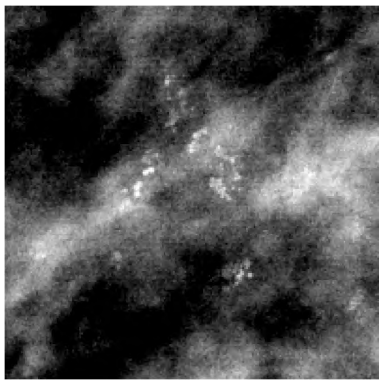

Figure 1.2: Example a cluster of microcalcifications

When tumors become invasive they can appear on mammograms as a mass or architectural distortion. If the mass is surrounded by a radiating pattern of spicules, it is called a stellate lesion. Not all tumors have a central mass, especially lobular carcinomas are often only detectable due to an architectural distortion of the breast tissue. In Figure 1.3 typical examples of a mass, a stellate lesion and an architectural distortion are shown. However, in practice a whole spectrum of appearances from lesions without a central mass, lesions with both a mass and spicules to lesions without any spiculation is found.

Sometimes a mass can only be identified as a tumor because an asymmetry is present between the left and right breast. Mammograms of both breasts should be more or less identical, a density present in only one image is suspect. To increase the sensitivity and specificity, mammograms from previous screening rounds are used to detect changes between the old and new films.

Some benign processes yield lesions that are hard to discriminate from malignant lesions. However, when a lesion has spicules or a faint jagged edge, it is likely to be malignant. When the edge of the lesion is sharp and well-described, it is more likely to be benign. Often, masses and microcalcifications occur together in one mammogram, making detection and classification easier.

1.3 Screening for breast cancer

In many countries a breast cancer screening program using mammography has been started to detect cancers as early as possible. A screening program is defined as a program where

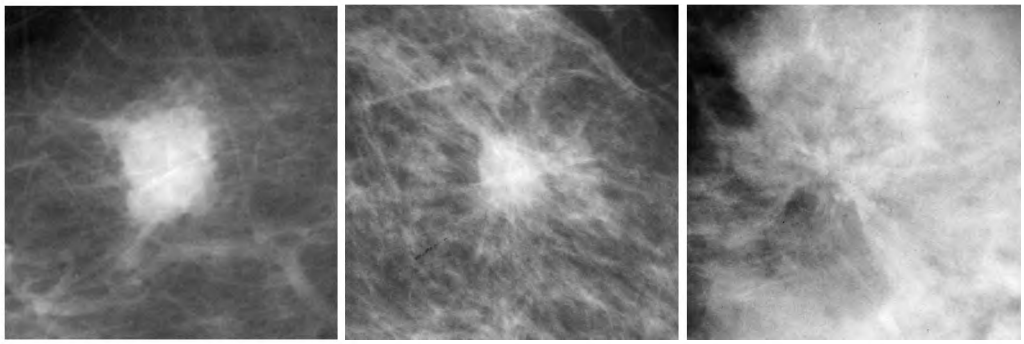

Figure 1.3: Examples of the most common signs of malignant abnormalities. Left: a circumscribed lesion. Middle: a stellate lesion. Right: an architectural distortion.

an asymptomic group is invited for examination for a specific disease on a regular basis. For breast cancer screening programs, only women are invited due to the very low incidence rate among men. A number of parameters must be chosen for a breast cancer screening program. The two main parameters are the age range of women that are invited and the time interval between two screening rounds. It is a highly debated subject at what age women should be invited for their first screening [49], varying in practice between 40 and 50 years. Below the age of 40, the incidence rate of breast cancer is extremely small, increasing rapidly between the age of 40 and 50, and continues to increase more gradually for older women. The problem with screening young women is that their breasts contain much glandular tissue, yielding mammograms that are difficult to read due to dense tissue. Breast cancers in young women are often aggressive, fast growing tumors, requiring short intervals between two screens. After menopause, the breast becomes less dense, making successful screening for small cancers more feasible. The upper limit of age for which women are invited for screening varies between 65 and 75 [17].

If the interval between two successive screening rounds is too large, a number of tumors that are detected in screening have already reached a stage with a low chance of successful treatment. Also, a number of tumors will occur during this interval, which are called interval carcinomas. A large number of interval carcinomas may indicate that the screening interval should be made shorter. A short interval period will have a larger effect on the reduction of mortality, but is more expensive and women are exposed to a higher number of X-ray doses. In the UK, the screening interval is 3 years, a period that is considered too long by some researchers [16], in Sweden and the Netherlands it is 2 years. Sometimes, the interval is made age dependent, 1 to 1.5 years for women below 50, 2 years after 50 [63].

In the UK only oblique films are used in screening, most other countries use both oblique and cranio-caudal films. The way mammograms are read also varies between countries. In some countries (for example the Netherlands) mammograms are examined by two radiologist, called double reading. Various approaches can be used to combine the findings of the two radiologists. Thurfjell has shown that the sensitivity increases when double reading is practiced when a case is recalled if either one of the radiologists finds it suspicious [65], while the positive predictive value did not change. This study was criticized [4] because the demonstrated increase in sensitivity is a mathematical fact, the question should be if the increase in sensitivity is worth the decrease in specificity. Another study has shown that double reading based on consensus between the radiologists is a cost-effective screening

procedure [8].

Proving the efficacy of a screening program in a traditional epidemiological way is difficult due to the lack of an effective control group [16]. If half of the population is offered screening, part of this group will not participate. Women in this group that develop breast cancer typically do not seek medical attention until tumors are already in a late and incurable stage. If the group of non participating women is large, a serious self-selection bias is present in the study. Even more important, women in the control group cannot be denied to have mammography on a regular basis. Especially women that are in a high risk group will do this, reducing the number of cancers that are found at an incurable stage in the control group. This effect is called contamination and is a serious problem when the effect of screening is studied. These two factors make it difficult to prove a significant mortality reduction in screening. Comparing the number of breast cancer deaths with the number before screening was started is a common way to solve this problem but suffers from several drawbacks: the incidence of breast cancer may have changed, the treatment of breast cancer may have improved, or women may be more aware of abnormalities and seek medical assistance earlier than they might have done before. Another complicating factor is the long time it takes before a screening program reaches its maximal reduction in mortality.

Early screening programs were based on breast examination using palpation, either by the woman or a physician. No significant reduction in mortality has been reported on randomized trials using this type of screening [44], although a few other studies suggest a small benefit [2]. In 1963 the HIP project was started in New York, the first large screening experiment using mammography as the main screening modality, together with palpation. A reduction in mortality was found for women in the group that underwent screening, a success that could be achieved because palpation and mammography were hardly practiced by women in the control group [58]. The success that was reported stimulated other countries like Sweden, Finland, the United Kingdom, Canada and the Netherlands to start experiments with breast cancer screening. Results of a number of large randomized studies have been published (see the overview in [16]). Many of these studies suffered from refusers and contamination, especially the Canadian studies have been criticized for their bias. However, it is commonly accepted that these studies show a reduction in mortality for women that take part in a screening program, especially for the age group between 50 and 70 years old. This is confirmed by other non-randomized and cohort studies in the UK and the Netherlands.

In a number of studies the possible benefit of inviting young women between 40 and 50 years old to a screening program was examined, but no unequivocal results were obtained. Some studies suggest a reduction in mortality [66], others do not find evidence for this [49]. It was shown by van Dijck et al [17] that screening is beneficial at least until the age of 75. Due to the limited number of women over 75 that participated in screening, no significant results could be obtained for this age group.

It is important that the radiographers are dedicated to their work because constant high technical quality is required to make a screening program successful. Radiologists should be motivated and experienced in reading mammograms [56, 16]. Studies have shown that there are large differences in performance between radiologists [57, 18, 3], only experienced and dedicated radiologists should work in breast cancer screening programs. Radiologists should receive the histological results of cases they recalled to increase their knowledge and to be aware of changes in their recall policy [16, 38]. Periodic audits should be held so that radiologists get feedback on their false negative interpretations [11].

1.4 Screening in the Netherlands

Since 1996, breast cancer screening in the Netherlands is nation wide. All women between 50 and 70 years are invited every two years to have a breast examination using mammography. In 1998 the upper limit was raised to 75 years. If a woman participates for the first time, both oblique and cranio-caudal films are made. On successive visits only oblique films are made, cranio-caudal films are added in approximately 20% of the cases when the radiographers finds the mammograms hard to read due to the presence of dense tissue or if a possible abnormality is visible. All mammograms are read by two radiologists, but not in a blinded way. When the second reader examines the films, the examination report of the first reader is available. If the two radiologist disagree on a specific case, they discuss the case and come to a consensus.

One of the targets in the Dutch screening program is a low number of unnecessary recalls. Between 1990 and 1997, of the women that participated for the first time in the screening, on average 13.1 out of 1000 were recalled for further examination [19]. Of these women, 6.1 were diagnosed with breast cancer after a follow-up examination. For women in successive screening rounds, 6.9 out of 1000 were recalled, and 3.5 were diagnosed with breast cancer. The total recall rate in the screening in 1997 was 9.3 out of 1000, yielding 4.4 cancers per 1000 screened women. This number is much higher than in other countries. In the United States and the UK the percentage of women that is recalled is between 5 to 10%. The number of women that was recalled has decreased over the years, for subsequent screening rounds from 1.22% in 1990 to 0.84% in 1996, but increased for the first time in 1997 to 0.93%. The additional tumors that are found when the recall rate is increased slightly may have a strong effect on the mortality reduction because these tumors are often in an early stage of development.

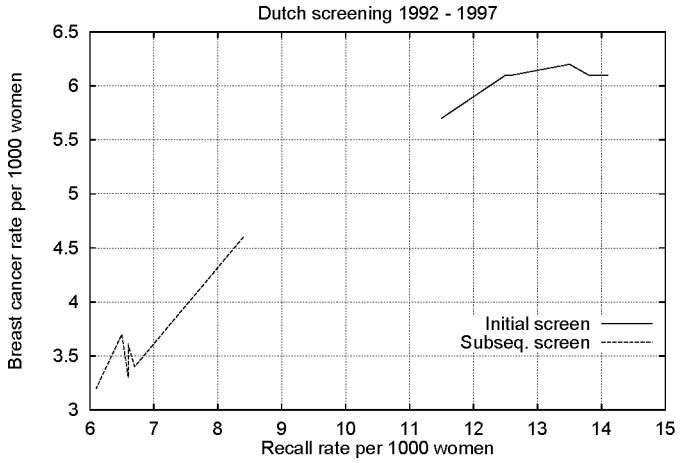

Figure 1.4: Tumor detection rate as a function of the recall rate for initial screens and subsequent screens.

Moskowitz suggests that a causal relation exists between the decreasing recall rate practiced in the Dutch screening and the increasing number of interval carcinomas that are found [40]. In Figure 1.4, the tumor detection rate is plotted as a function of the recall rate for the initial screens and the subsequent screens. A relation between the recall rate and the detection rate is shown, suggesting that when the specificity is lowered the detection rate will

increase. Note that the strong relation that is shown for the "all screens" set is for a big part due to the shift of the majority of the women from the initial screen group to the subsequent screen group during these years. In 1992, approximately 75% of the screens were initial screens, in 1997 only 25%.

Attendance and compliance are high in the Dutch screening. In 1997, the attendance was almost 80% of all invited women, a number that is constant over the 4 screening rounds that were held by 1997 [19]. The high attendance and compliance may be a result of the restrictive recall policy that is practiced.

The most recent figures suggest a mortality reduction since screening has started in the Netherlands, but no significant results are found yet. It is expected that significant reduction of breast cancer related deaths will be found for the first time next year [19].

1.5 Errors in screening

Several studies have shown that approximately 20% of all interval carcinomas was visible on a previous screening mammogram [55, 69, 9]. Of all screening detected cancers, also 20% is retrospectively considered actionable on a previous screening mammogram [25, 5, 67]. These numbers suggest that a considerable improvement in mortality reduction is possible if these errors could be prevented. When mammograms are examined retrospectively for signs of malignancy, the abnormality is considered occult (nothing visible on the previous mammogram), or either classified as minimal sign or a screening error. An abnormality is called a minimal sign if something abnormal is found in the region of interest that is not suspicious enough to recall. If signs of malignancy are present that are actionable, it is called a screening error. It is likely that different standards are used when a mammogram is examined retrospectively for visible signs of malignancy because the radiologist knows a tumor was found within the next 2 years. This was confirmed by a study by Harvey [25], who showed that retrospective reviews do not reflect the everyday reading practice. However, many tumors do show clear signs of malignancy on previous mammograms, and many are found by an automated detection system at high specificity levels. A problem with this type of studies is the subjective nature of the classes normal, minimal sign and screening error, since the used definitions vary considerably between radiologists.

There can be two reasons why women with a visible tumor were not recalled for follow-up research. The first possibility is that the sign was overlooked, and has not been examined at all by the radiologist. The second possibility is that the sign was examined but was considered benign, normal, or not found suspicious enough for further examination. Kundel and Nodine [36] defined three types of error in detection of lung nodules: search, recognition and decision. If the fovial search area did not pass the region with the abnormality it was called a search error. If the area containing the abnormality was in the fovial scan path, a difference is made between two type of errors: recognition and decision. If the dwell time at the location of the abnormality was longer than 0.3 seconds, it was called a decision error. If the time was shorter than 0.3 seconds, it was called a recognition error. This debatable threshold level suggests that a grey area is present between "examined" and "not examined". Krupinski investigated gaze duration and location, and found that in mammographic search on average 87% of the breast was examined [35]. Areas with false negatives were examined for a longer period than true negative areas, but shorter than true and false positive

areas. So far, only a few studies have focused on the reasons why errors are made in the field of mammography [24, 28], although some work has been done in other medical areas [20]. Much work on signal detection theory [22] has been done in psychology departments, some related to the medical field [38]. Psychophysical evidence exists that inserting extra abnormal signals to increase the target rate improves the performance when the target rate is very low, which is the case in breast cancer screening programs [38]. Insertion of benign and malignant abnormalities to increase the rate of abnormalities in a screening program would have the additional benefit that easy control of the quality of the performance of the radiologist is possible.

The mammographic signs missed most in screening programs are masses and architectural distortions [5, 9, 69]. Masses are often obscured by glandular tissue, or have low contrast or no clear malignant signs, like a fuzzy edge or spicules. Microcalcifications are more easily detected by radiologists, but are often hard to classify in benign or malignant types.

1.6 Computer aided diagnosis

An important development that may help to improve the performance in breast cancer screening as well as clinical practice is computer aided diagnosis (CAD). Mammograms have to be digitized before automated methods can search them for abnormalities, but in the next 3-5 years, digital mammography systems will enter the clinics and screening centra, making CAD feasible. The last 10 years, much research has been done in the field of digital mammography, mainly in the UK and the United States. Computers can be used for many tasks in breast cancer screening programs. Mammograms can be enhanced for optimal viewing conditions, software can help searching for suspicious signs, or could help classifying lesions or microcalcifications in benign or malignant types.

Most work in CAD has focused on preventing search errors by prompting suspiciously looking regions. If most lesions are prompted, the number of errors due to oversight will be diminished. High sensitivity, however, comes at the cost of low specificity, which will result in a low positive predictive value (PPV) per prompt. Hutt [29] has shown that prompting systems work if the number of false positive prompts is sufficiently low. A number of experiments in mammography have shown that radiologists that are aided by a prompting system work at a higher performance level than a radiologist working without prompts [13, 34]. Prompting systems can be used at a range of different sensitivity/specificity levels. It is still an open question at which settings the reduction in screening errors will be maximized, a setting that might vary for radiologists. When the operation point is chosen such that high sensitivity is achieved, the PPV will become low and radiologists may not take prompts serious. If the operation point is chosen at a high specificity point, the PPV will be much higher, but not all tumors will be signaled.

This section describes the various tasks CAD can be used for, and gives references to some important approaches and algorithms.

1.6.1 Image processing

Many papers have been published on enhancing mammograms for optimal viewing, mainly for detection of microcalcifications [41, 60, 12]. However, most of these studies do present nice images but do not provide evidence that radiologists perform better on these processed images. Important work was done to transform the mammogram in such away that it can be printed or examined on a monitor optimally [1, 43]. For example, the dark area near the skin line can be enhanced [33, 10] and the pectoral muscle can be filtered out, largely reducing the intensity range in the mammogram. Good contrast will be available in the whole area of interest, both in the pectoral area as well as near the skin line.

Highnam et al. [26] described a method to filter scatter from the mammogram. The anti-scatter grid may not be required anymore, making a large reduction in the X-ray dose possible.

1.6.2 Detection and classification of microcalcifications

Dealing with noise in mammograms is very important for microcalcification detection algorithms. Most methods use a form of local adaptive thresholding, because noise levels vary across the image. Nishikawa et al. used an initial global threshold level, followed by a locally adaptive threshold step [46]. Chitre et al. [15] computed a local threshold image and used the local deviation of grey levels as a threshold to decide whether or not a pixel belonged to a microcalcification cluster. If the noise is made independent of the signal, global threshold values can be used on the local contrast images. Karssemeijer developed a method for this purpose [32], which was improved by Veldkamp et al. [68]. Advantage of a global correction approach is that the statistics are much better than for local estimations of the appropriate threshold level. Local structures like lines can have a disturbing effect on the local estimation of the threshold level. Several other groups focused on noise equalization as well [42, 61].

Classification of calcification clusters is an important topic in CAD because it is a task radiologist find hard. To achieve a good positive predictive value (PPV) it is important to be able to discriminate between malignant and benign microcalcification clusters, because only 20% of all clusters are due to malignant processes. Jiang et al. developed a method that outperformed five radiologists [31], using a neural network that classified clusters based on eight features that were computed for each cluster. The clusters were manually identified, but the areas of the calcifications were grown automatically. Chan [14] used a similar approach to construct a cluster to classify, because their detection algorithm did not find all clusters and included many false positive calcifications. This manual step might induce a bias, because for malignant clusters the radiologist might point out the the calcifications that make the cluster look malignant, and in a subconscious way can enter his knowledge in the annotation. Also, pointing out all calcifications is a tedious and time consuming task, not suited for use in clinical practice. So far only few completely automated methods have been published [59].

1.6.3 Detection and classification of masses and architectural distortions

A large variety of techniques have been applied to the problem of mass detection, but most follow a two-step scheme that was described by Woods and Bowyer [71]. First, one or more features are computed for each pixel, after which each pixel is classified and the suspicious pixels are grouped into a number of suspicious regions. In a second step, these regions are classified as normal or abnormal regions, based on regional features like size, shape or contrast.

Two signs can indicate the presence of a lesion: a radiating pattern of spicules or a central mass. To detect the whole range from architectural distortions to circumscribed masses, both signs much be detected.

Detection of the central mass

The central mass is a more or less circular bright region with a diameter between 5 mm and 5 cm. Convolution of the image with a zero-mean filter with a positive center and a negative surrounding area was used by a number of research groups to detect the mass [53, 75], for example with the Laplacian of the Gaussian (LoG) or a Difference of Gaussians filter (DoG). This is an easy and intuitive approach to detect bright blobs, but may not be suited to find masses with low contrast. Other approaches that are less dependent of the contrast may be more useful, like template-matching, a method used in some of the earlier papers in this field [45, 37]. A model is made of the appearance of a mass, and the mammogram is searched for regions that resemble this model. This approach is more related to the shape, and less to the contrast of the region. Especially for hard to detect low contrast masses this method may outperform convolution based approaches.

Most recent methods for mass detection focus on the analysis of the gradient patterns in an area of interest. The appearance of masses in mammograms varies and therefore the above described rigid approaches are not very successful. In an area with a central mass, the orientation of the gradients will be towards the center of the mass. Statistical analysis of this pattern can be used to discriminate masses from other structures. Groshong and Kegelmeyer [23] used a generalized Hough transform for circles. The strongest edges in an area of interest are accumulated in a Hough space where each location relates to a center and a radius. Masses will yield peaks in this space. Zwiggelaar et al. applied a one-dimensional recursive median filter over a number of different angles to each pixel [76]. Based on the variations in scale for various angles they can determine whether the structure is a blob or has a more linear shape.

Sometimes a mass looks very much like normal glandular structure, and is only detectable due to asymmetry between the left and right breasts. A few papers have been published describing approaches for mass detection based on differences in left and right mammograms. These approaches perform some kind of image subtraction, and can also be used to detect temporal changes when a mammogram is compared with an older mammogram of the same breast. Matching two breasts is a complicated procedure because there is only an approximate correspondence between the normal tissue in the two breast, and due to variations in compression and positioning the variation in appearance is even made larger. Yin and Giger [72, 73] applied a simple rigid body transform to align the skin line

of the two breasts. Other more sophisticated approaches match corresponding points in the two breasts. Lau and Bisschof [39] use a set of 3 control points and an estimation of the nipple, Sallam and Bowyer use a more general warping method to match automatically detected landmarks of the glandular tissue [54]. When the two breasts are correctly matched, subtraction and smoothing can be done to find a number of suspect regions [39], or a more advanced non-linear subtraction can be used [72, 73].

Detection of the spicules

When a mass is surrounded by spicules, it is likely to be malignant. Many stellate lesions are easier to detect by their spicules than by their central mass, and for architectural distortions it is the only sign. Kegelmeyer [34] computed histograms of local gradient orientations. Areas with a spicule pattern should have flatter histograms than normal areas. This feature was combined with four other texture features, but his very good results could not be reproduced by other groups [71]. Analysis of texture in the Hough space was the basis of a method published by Zhang and Giger [74]. Parr et al. [47, 48] developed a model for spicules using principal component analysis.

False positive removal: classifying regions as normal or abnormal

Pixel-level detection algorithms that detect both spicules and masses can be very sensitive, but will signal many false positives. A number of papers have been published on the topic of discriminating real lesions from suspiciously looking normal tissue [70, 53, 51, 73]. Sometimes, texture features are computed over a large region containing the suspicious region [70], but most groups segment the suspicious area. Segmentation of the suspicious area is useful in separating abnormal and normal tissue, because it enables computation of features related to the edge of the region, as well as contrast and shape features. A considerable improvement is generally achieved when region based features are computed to remove false positive signals.

Classification of lesions to benign and malignant types

Classification of benign and malignant masses is a well studied subject in mammography. All papers focus on edge analysis of the mass, a vague or spiculated edge indicates malignancy, a sharp well defined contour is likely to belong to a benign abnormality. Other features that are computed in many papers are size, shape, texture and contrast measures. In some work the lesions were outlined by the radiologist [52], a time consuming task which can induce a bias because the way a radiologist outlines spicules or vague regions may be incorporated in the computed features. Therefore, automatic segmentation is preferable. Just like for classification of microcalcifications, a performance comparable to radiologists is achieved nowadays. Interesting work was done by Giger and Huo [21, 27] who used a radial edge gradient method to discriminate malignant and benign lesions, and Pohlman et al [50] who developed a feature describing tumor boundary roughness.

1.7 Outline of this thesis

This topic of thesis is automated detection of masses and distortions in mammograms. Detection of spicules and the central mass is addressed, as well as region-level false positive removal that was described in the previous section.

Chapter 2 describes a method to detect the radiating pattern of spicules that can be found with architectural distortions and stellate lesions. A sensitive spicule detection method based on statistical analysis of second order line estimations is applied to a small set of architectural distortions and stellate lesions.

The statistical analysis that is described in Chapter 2 to detect spicules can also be used for mass detection. Instead of a second order line estimations to determine line like structures, first order gradient estimates can be computed to detect masses. Mass features were combined with spicule features to improve detection performance on the set of stellate lesions and architectural distortions, but because of a decrease in sensitivity on the architectural distortions no real improvement was found [64]. In Chapter 3 the method is applied to a large consecutive set of masses, stellate lesions and distortions taken from the Nijmegen screening program.

In many papers on mass detection, the method is applied in a multi-scale way, "because masses vary in size", but in few the benefit of using multi-scale over single-scale detection is shown. In Chapter 4, three different mass detection methods are applied both in a single and in a multi-scale way to examine the possible gain.

Mass detection algorithms achieve high sensitivity, but signal many false positives as well. To achieve high sensitivity and specificity, regional information like shape and contrast must be incorporated in the method to discriminate real masses from suspiciously looking normal tissue. Most false positive removal steps segment the suspicious region. Chapter 5 describes a mass segmentation method using a discrete dynamic contour model, and compares this method to a region growing approach. The segmented areas are compared to the annotations made by the radiologist using an overlap criterion.

Chapter 6 describes a false positive removal procedure that can follow a mass detection algorithm that has detected suspicious regions in mammograms. The method is based on the mass segmentation method that is described in Chapter 5. Using the segmentation, a number of features are computed that are based on decision criteria that are also used by radiologists to discriminate real lesions from normal tissue. It is expected that a significant reduction in false positives will be achieved.

A data set was constructed with mammograms containing signs of malignancies that were only detected in screening two years later, or became interval carcinomas. If these false negative cases can be detected by an automated detection system at reasonable specificity levels, it is an indication that CAD can be useful tool for radiologists. Chapter 8 describes the experiment using the detection method described in Chapter 3.

A commercially available CAD system was placed in Nijmegen, which enabled us to do some small studies on how radiologist work with such a system. Chapter 9 describes two studies that were carried out using this system. The main focus of the first study was the specificity of radiologist using a CAD system; some fear it will increase the number of false positive recalls. The inter-observer variation was examined, as well as the type of abnormalities that cause problems for radiologists. In the second study, radiologists were asked to mark and grade all regions they found suspicious, to shed light on the type of

regions that radiologists find difficult to interpret.

Bibliography

[1] SR Aylward, BM Hemminger, ED Pisano, and RE Johnston. Mixture modeling for digital mammogram display and analysis. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 305–312. Kluwer, Dordrecht, 1998.

[2] C J Baines. Breast self-examination. *Cancer*, 69(7 suppl):1942–1946, 1992.

[3] C A Beam, P M Layde, and D C Sullivan. Variability in the interpretation of screening mammograms by u.s. radiologists. *Arch Intern Med*, 156:209–213, 1996.

[4] C A Beam and D C Sullivan. What are the issues in the double reading of mammograms? *Radiology*, 193(2):582, Nov 1994. Letter.

[5] R E Bird, T W Wallace, and B C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184:613–617, 1992.

[6] C Boetes, J O Barentsz, R D Mus, R F van der Sluis, L J van Erning, J H Hendriks, R Holland, and S H Ruys. Mr characterization of suspicious breast lesions with a gadolinium-enhanced turboflash subtraction technique. *Radiology*, 193(3):777–781, 1994.

[7] C Boetes, R D Mus, R Holland, J O Barentsz, S P Strijk, T Wobbes, J H Hendriks, and S H Ruys. Breast tumors: comparative accuracy of mr imaging relative to mammography and us for demonstrating extent. *Radiology*, 197(3):743–747, 1995.

[8] J Brown, S Bryan, and R Warren. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ*, 312(7034):809–812, 1996.

[9] H C Burrel, D M Sibbering, A R M Wilson, S E Pinder, A J Evans, L J Yeoman, C W Elston, I O Ellis, R W Blamey, and J F R Robertson. Screening interval breast cancers: mammographic features and prognostic factors. *Radiology*, 199:811–817, 1996.

[10] J W Byng, J P Critten, and M J Yaffe. Thickness-equalization processing for mammographic images. *Radiol*, 203:564–568, 1997.

[11] G Cardenosa. Mammography: An overview. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 3–10. Elsevier, Amsterdam, 1996.

[12] E Cernadas, L Gomez, P G Rodriguez, A Casas, R G Carrion, and J J Vidal. Design of unsharp masking filters in the frequency domain: Parametrization for breast radiographs. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 463–466. Elsevier, Amsterdam, 1996.

[13] H P Chan, K Doi, C J Vyborny, R A Schmidt, C E Metz, K L Lam, T Ogura, Y Wu, and H Macmahon. Improvement in radiologist's detection of clustered microcalcifications on mammograms. *Inv Radiol*, 25:1102–1110, 1990.

[14] H P Chan, B Sahinerand K L Lam, N Petrick, M A Helvie, M M Goodsitt, and D D Adler. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. *Med Physics*, 25(10):2007–2019, Oct 1998.

[15] Y Chitre, A P Dhawan, and M Moskowitz. Classification of mammographic microcalcifications using image structure and cluster features. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 31–40. Elsevier, Amsterdam, 1994.

[16] P B Dean. Overview of breast cancer screening. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 19–26. Elsevier, Amsterdam, 1996.

[17] J A Van Dijck, A L Verbeek, L V Beex, J H C L Hendriks, R Holland, M Mravunac, H Straatman, and J M Werre. Breast-cancer mortality in a non-randomized trial on mammographic screening in women over age 65. *Int J Cancer*, 70(2):164–168, 1997.

[18] J G Elmore, C K Wells, C H Lee, D H Howard, and A R Feinstein. Variability in radiologists' interpretations of mammograms. *N Engl J Med*, 331(22):1493–1499, 1994.

[19] Landelijk evalutatie team voor bevolkingsonderzoek naar borstkanke. Landelijke evaluatie van bevolkingsonderzoek naar borstkanke. Instituut voor maatschappelijke gezondheidszorg, Erasmus universiteit Rotterdam, Postbus 1738, 3000 DR, Rotterdam, 1997. ISBN 90-72245-89-X.

[20] P J Friedman. The past and future of radiological error. In E A Krupinski, editor, *Medical imaging 1999: Image perception and performance*, volume 3663, pages 2–7, 1999.

[21] M L Giger, C J Vyborny, and R A Schmidt. Computerized characterization of mammographic masses : analysis of spiculation. *Cancer Letters*, 77:201–211, 1994.

[22] D M Green and J A Swets. *Signal detection theory*. Wiley, New York, 1966.

[23] B R Groshong and W P Kegelmeyer. Evaluation of a hough transform method for circumscribed lesion detection. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 361–366. Elsevier, Amsterdam, 1996.

[24] M Hartswood, R Procter, and LJ Williams. Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 363–370. Kluwer, Dordrecht, 1998.

[25] J E Harvey, L L Fajardo, and C A Inis. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. *AJR*, 161:1167–1172, 1993.

[26] R P Highnam, J M Brady, and B J Shepstone. Removing the anti-scatter grid in mammography. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 459–462. Elsevier, Amsterdam, 1996.

[27] Z Huo, M L Giger, C J Vyborny, U Bick, P Lu, D E Wolverton, and R A Schmidt. Analysis of spiculation in the computerized classification of mammographic masses. *Med Phys*, 22:1569–1579, 10 1995.

[28] I Hutt. *The computer-aided detection of abnormalities in digital mammograms*. PhD thesis, University of Manchester, Faculty of Medicine, Department of Medical Biophysics, 1996.

[29] I W Hutt, S M Astley, and C R M Boggis. Prompting as an aid to diagnosis in mammography. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 389–398. Elsevier, Amsterdam, 1994.

[30] V P Jackson. The role of us in breast imaging. *Radiology*, 177(2):305–311, 1990.

[31] Y Jiang, R M Nishikawa, D E Wolverton, C E Metz, M L Giger, R A Schmidt, C J Vyborny, and K Doi. Malignant and benign clustered microcalcifications: automated feature analysis and classification. *Radiology*, 198(3):671–678, Mar 1996.

[32] N Karssemeijer. A stochastic model for automated detection of calcifications in digital mammograms. *Image and Vision Computing*, 10:369–375, 1992.

[33] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[34] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[35] E A Krupinski and C F Nodine. Gaze duration predicts the locations of missed lesions in mammography. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 399–405. Elsevier, Amsterdam, 1994.

[36] H L Kundel and C F Nodine. Studies of eye movements and visual search in radiology. In J A W Seders, D Fisher, and R Monty, editors, *Eye movements and the higher psychological functions*. Hillsdale NJ, 1978.

[37] S M Lai, X Li, and W F Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans on Med Imag*, 8:377–386, 1989.

[38] D Laming. Screening cervival smears. *British Journal of Psychology*, 86:507–516, 1995.

[39] T K Lau and W F Bischof. Automated detection of breast tumors using the asymmetry approach. *Comp and Biomed Research*, 24:273–295, 1991.

[40] M Moskowitz. Retrospective reviews of breast cancer screening: what do we really learn from them? *Radiology*, 199(3), 1996.

[41] R Mutihac, AA Colavita, A Cicuttin, and A Cerdeira. Maximum entropy improvement of x-ray digital mammograms. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 329–336. Kluwer, Dordrecht, 1998.

[42] T Netsch. *Detection of microcalcification clusters in digitized mammograms*. PhD thesis, University of Bremen, 1998.

[43] T Netsch, M Biel, and HO Peitgen. Display of high-resolution digital mammograms on crt monitors. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 313–320. Kluwer, Dordrecht, 1998.

[44] P A Newcomb, N S Weiss, B E Storer, D Scholes, B E Young, and L F Voigt. Breast self-examination in relation to the occurrence of advanced breast cancer. *J Nath Cancer Inst*, 83(4):260–265, 1991.

[45] S L Ng and W F Bischof. Automated detection and classification of breast tumors. *Comput Biomed Res*, 25:218–237, 1992.

[46] R M Nishikawa, M L Giger, K Doi, C J Vyborny, and R A Schmidt. Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms. In K W Bowyer and S M Astley, editors, *State of the art in digital mammographic image analysis*, volume 9 of *Series in machine perception and artificial intelligence*, pages 82–102. World Scientific, 1994.

[47] T C Parr, S M Astley, C J Taylor, and Boggis CRM. Model based classification of linear structures in digital mammograms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 351–356. Elsevier, Amsterdam, 1996.

[48] T C Parr, C J Taylor, S M Astley, and Boggis CRM. A statistical representation of pattern structure for digital mammography". In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 357–360. Elsevier, Amsterdam, 1996.

[49] P G M Peer, J M Werre, M Mravunac, J H C L Hendriks, R Holland, and A L M Verbeek. Effect on breast cancer mortality of biennial mammographic screening of women under age 50. *Int J Cancer*, 60:808–811, 1995.

[50] S Pohlman, K A Powell, N A Obuchowski, W A Chilcote, and S Grundfest-Broniatowski. Quantitative classification of breast tumors in digitized mammograms. *Med Phys*, 23:1337–1345, 1996.

[51] W E Polakowski, D A Cournoyer, S K Rogers, M P DeSimio, D W Ruck, J W Hoffmeister, and R A Raines. Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency. *IEEE transcations on medical imaging*, 16(6):811–819, December 1997.

[52] R M Rangayyan, N M El-Faramawy, J E L Desautels, and O A Alim. Measures of acutance and shape for classification of breast tumors. *IEEE transactions on medical imaging*, 16(6):799–810, december 1997.

[53] B Sahiner, H P Chan, N Petrick, D Wei, M A Helvie, D D Adler, and M M Goodsitt. Classification of mass and normal breast tissue : a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imag*, 15:598–610, 10 1996.

[54] M Sallam and K W Bowyer. Registering time-sequences of mammograms using a two-dimensional unwarping technique. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 121–131. Elsevier, Amsterdam, 1994.

[55] C J Savage, A G Gale, E F Pawley, and A R M Wilson. To err is human; to compute divine? In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 405–414. Elsevier, Amsterdam, 1994.

[56] F Schmidt, K A Hartwagner, E B Spork, and R Groell. Medical audit after 26,711 breast imaging studies: improved rate of detection of small breast carcinomas (classified as tis or t1a,b). *Cancer*, 83(12):2516–2520, 1998.

[57] R A Schmidt, R M Nishikawa, R B Osnis, K L Schreibman, M L Giger, and K Doi. Computerized detection of lesions missed by mammography. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 105–110. Elsevier, Amsterdam, 1996.

[58] S Shapiro, P Strax, and L Venet. Periodic breast cancer screening in reducing mortality from breast cancer. *JAMA*, 215(11):1777–1785, 1971.

[59] E Sorantin, F Schmidt, H Mayer, P Winkler, C Szepesvari, E Graif, and E Schuetz. Automated detection and classification of microcalcifications in mammograms using artificial neural nets. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 225–232. Kluwer, Dordrecht, 1998.

[60] R N Strickland, L J Baig, W J Dallas, and E A Krupinski. Wavelet-based image enhancement as an instrument for viewing cad data. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 441–446. Elsevier, Amsterdam, 1996.

[61] R N Strickland and H I Hahn. Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Trans. on Medical Imaging*, 15(2):321–332, April 1996.

[62] L Tabar and P Dean. *Teaching atlas of mammography*. Georg Thieme Verlag, New York, 1985.

[63] L Tabar, G Fagerberg, N E Day, and L Holmberg. What is the optimum interval between mammographic screening examinations? an analysis based on the latest results of the swedish two-county breast cancer screening trial. *Br J Cancer*, 55(5):5470551, 1987.

[64] G M te Brake and N Karssemeijer. Detection of stellate breast abnormalities. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 341–346. Elsevier, Amsterdam, 1996.

[65] E L Thurfjell, K A Lernevall, and A A S Taube. Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191:241–244, 1994.

[66] E L Thurfjell and J A Lindgren. Breast cancer survival rates with mammographic screening: similar favorable survival rates for women younger and those older than 50 years. *Radiology*, 201(2):421–426, 1996.

[67] J A M van Dijck, L M Verbeek, Hendriks J H C L, and R Holland. The current detectability of breast cancer in a mammographic screening program. *Cancer*, 72:1933–1938, 1993.

[68] W Veldkamp and N Karssemeijer. Improved correction for signal dependent noise applied to automatic detection of microcalcifications. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 169–176. Kluwer, Dordrecht, 1998.

[69] B Vitak. Invasive interval cancers in the Östergötland mammographic screening programme: Radiological analysis. *European Radiology*, 8:639–646, 1998.

[70] D Wei, H P Chan, M A Helvie, B Sahiner, N Petrick, D D Adler, and M M Goodsitt. Classification of mass and normal breast tissue on digital mammograms : multiresolution texture analysis. *Med Phys*, 22:1501–1513, 9 1995.

[71] K Woods and K Bowyer. A general view of detection algorithms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 385–390. Elsevier, Amsterdam, 1996.

[72] F F Yin, M L Giger, K Doi, C E Metz, C J Vyborny, and R A Schmidt. Computerized detection of masses in digital mammograms : Analysis of bilateral substraction images. *Med Phys*, 18:955–963, 1991.

[73] F F Yin, M L Giger, C J Vyborny, K Doi, and R A Schmidt. Comparison of bilateral-substraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol*, 6:473–481, 1993.

[74] M Zhang and M L Giger. Automated detection of spiculated lesions and architectural distortions in digitized mammograms. *SPIE 2434*, pages 846–855, 1995.

[75] B Zheng, Y H Chang, and D Gur. Computerized detection of masses in digitized mammograms using single image segmentation and a multi-layer topographic feature analysis. *Acad Radiol*, 2:959–966, 1995.

[76] R Zwiggelaar, T C Parr, J E Schumm, I W Hutt, C J Taylor, S M Astley, and CRM Boggis. Model-based detection of spiculated lesions in mammography. *Medical Image Analysis*, 3(1):39–62, 1999.

Chapter 2

Detection of stellate distortions in mammograms[1]

2.1 Introduction

In breast cancer screening, detection of malignant abnormalities must be performed at a very high level of specificity, because of the large number of normal cases. For instance, in the Dutch nationwide screening program, in about four out of a thousand screening cases a cancer is found after the first screening round, while a positive predictive value of 70 percent is aimed at. In addition, mammographic signs indicating early stages of breast cancer can be very subtle, making screening a difficult task for radiologists. The variation of normal parenchymal patterns is large, and benign lesions may be hard to distinguish from cancers. Because of these reasons, screening mammograms are often read independently by two radiologists to reduce false negatives and interpretation errors. In spite of this, analysis of previous screening mammograms of women with breast cancer still reveal a substantial fraction of radiological errors.

In the near future, the introduction of digital mammography systems will allow application of computerized methods for detecting subtle abnormalities in mammographic images. Pattern recognition techniques are being developed for this purpose, and have already been shown to have the potential to improve radiological performance [8]. In general, these methods tend to be very sensitive but less specific. Therefore, they seem to be well suited to mark suspicious mammographic areas in order to attract the radiologists attention. Apart from improving detection rates, such a computer-aided setup may remove the need for double reading.

Retrospective studies analyzing the types of carcinomas missed in breast cancer screening [14, 12] reveal that minimal signs on previous screening mammograms of patients with breast cancer are most often classified as vague densities or masses. Therefore, computer programs for detection of masses are likely to become important in breast cancer screening. They can be used as a tool for radiologists to attract their attention to suspicious areas. In general, malignant mammographic densities have an irregular appearance, often surrounded by a radiating pattern of linear spicules. Sometimes the density is very faint, and when it is

[1]Published as: N. Karssemeijer, G.M. te Brake *Detection of Stellate Distortions in Mammograms*, IEEE Transactions on Medical Imaging, vol 15, Nr. 5, 611-619, 1996

embedded in the parenchymal tissue it may be very hard to perceive. In those cases the stellate pattern of spicules is the most important sign. The aim of this investigation is detection of such stellate patterns without relying on the presence of a central mass. The approach is based on statistical analysis of a map of pixel orientations. The idea is that if an increase of pixels pointing to a given region is found then this region may be suspicious, especially if, viewed from the test region, such an increase is found in many directions. It is noted that no attempt is being made to identify spicules explicitly. In this respect the approach is similar to the texture based approach suggested by Kegelmeyer [7], and different from approaches based on using the Hough transform for spicule detection, and on accumulation of the evidence for spicules to detect tumors [11]. The latter approach relies on a separate stage for spicule detection and tends to be less successful in case spicules are very faint. Clearly, this is undesirable because cases with only faint signs are those in which computer-assisted reading will probably be the most helpful.

An important feature of the method is the way in which an orientation of the image intensity map is determined at each pixel. A new method is proposed for this purpose, based on the application of second order operators. If a line-like structure is present at a given site the method provides an estimate of the orientation of this structure, whereas in other cases the image noise will generate a random orientation. Using scale space theory it will be shown how accurate estimates of line orientation can be obtained at a given scale from the output of only three directional second-order Gaussian derivative operators, differing by $\pi/3$ in orientation.

The line-based orientation estimates are used to construct two operators which respond to radial patterns of straight lines. Combination of the output of these operators in a classifier leads to a very sensitive method for detection of stellate patterns. The method is applied to detect stellate lesions and architectural distortions in mammograms from the MIAS database, which was made available for public use by the Mammographic Image Analysis Society in the UK [13]. Earlier results of this project were reported in [4] and [5]. Since the initial report, the method has progressed considerably by improving feature calculation and implementation. This led to a significant increase in performance. For instance, an important improvement was obtained by adding a preprocessing step in which the decreasing tissue thickness near the breast skin line is compensated for. Another issue which is investigated in this paper is application of k-nearest neighbor, a neural network, and a decision tree for classification, in comparison to the Bayes classifier used in earlier work. Also the use of multiple scales for estimation of line orientation is validated by determining FROC performance at different single spatial scales.

2.2 A multi-scale line-based orientation map

Line orientation is often determined by using gradient operators. An important reason for this is the fact that two operators are sufficient to estimate edge orientation, while the result is relatively insensitive for changes of line-width. This makes the approach computationally attractive. If the image function is denoted by $I(x,y)$ the gradient of I in a given direction $r = (\cos(\phi), \sin(\phi))$ can be written as

$$\frac{\partial I}{\partial r} = \frac{\partial I}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial I}{\partial y}\frac{\partial y}{\partial r} = I_x \cos(\phi) + I_y \sin(\phi) \tag{2.1}$$

By maximization of this first-order directional derivative over ϕ, it follows that edge orientation can be computed by

$$\theta(x,y) = tan^{-1}\left(\frac{I_y}{I_x}\right) \tag{2.2}$$

Estimation of I_x and I_y can be performed by applying, for instance, the Sobel operator. For more accurate results a convolution with Gaussian derivatives should be performed [9, 2]. For continuous I it can be shown that this leads to homogeneous, isotropic and scale invariant results.

A drawback of the use of first order operators for estimation of line orientation lies in the fact that no valid results are obtained at the central part of the line. Moreover, the method is not sensitive for lines one pixel in width. These problems can be avoided by using second-order operators. Compass gradients have been suggested for this purpose, which roughly approximate second order directional derivatives $\partial^2 I/\partial r^2$ for lines of one pixel in width at a fixed set of angles [3]. Using this approach, line orientation can be estimated by selecting the filter with the highest output. Clearly, the accuracy of the estimates will depend on the number of angles at which the filters are being applied. Generalization to lines of different width leads to a multi-scale representation. An example of this approach using Gabor filters can be found in [1].

From a computational viewpoint, the use of a large filter bank is not very attractive. In particular, the use of the same neighborhood operator at many different angles seems inefficient. It can be shown by application of scale space theory that an accurate and more efficient method is possible. More precisely, at a given level of spatial scale σ, convolution of an image with only three filter kernels $W_\sigma(\theta_n)$ appears to be sufficient to make an accurate operator for determining line orientation, with $W_\sigma(\theta_n)$ the second order directional derivatives of the Gaussian kernel $G(r,\sigma)$ in the directions $\theta = n\pi/3$, $(n = 0,1,2)$, and

$$G(r,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{2.3}$$

The basis for this approach is a relation derived by Koenderink [9] which shows that for an arbitrary direction $W_\sigma(\theta)$ can be expressed as

$$\begin{aligned} W_\sigma(\theta) &= \frac{1}{3}(1+2\cos(2\theta))\,W_\sigma(0) \\ &\quad +\frac{1}{3}(1-\cos(2\theta)+\sqrt{3}\sin(2\theta))\,W_\sigma(\pi/3) \\ &\quad +\frac{1}{3}(1-\cos(2\theta)-\sqrt{3}\sin(2\theta))\,W_\sigma(2\pi/3) \end{aligned} \tag{2.4}$$

The three independent line operators W_σ form a non-orthogonal basis (Figure 2.1).

Using this relation we can calculate the filter orientation with maximum output by first solving $dW_\sigma(\theta)/d\theta = 0$, and then determining which of the extrema found is the maximum. Differentiation of W_σ leads to

$$\theta_{min,max} = \frac{1}{2}\left[\arctan\left(\sqrt{3}\frac{W_\sigma(2\pi/3)-W_\sigma(\pi/3)}{W_\sigma(\pi/3)+W_\sigma(2\pi/3)-2W_\sigma(0)}\right) \pm k\pi\right], \tag{2.5}$$

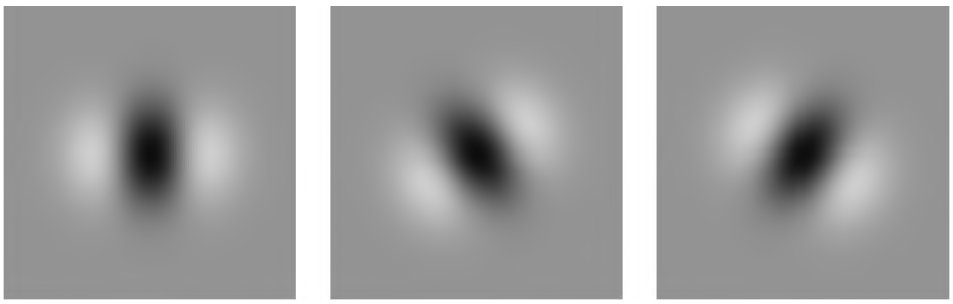

Figure 2.1: Three directional second order Gaussian derivatives used for estimation of line orientation

where one should note that θ gives the orientation perpendicular to the line itself. Thus, for a given site in the image two extrema are found at each scale. By using (4) the output of the filter at these orientations can be determined, and the orientation with maximum output can be determined. At this point, however, care must be taken because lines with positive contrast will give a strong negative output whereas lines with negative contrast give a positive output. Moreover, each scale will generate its own optimal orientation. Here, to condense the multi-scale filter output into one orientation per pixel, the orientation with the maximum absolute value is taken, at the scale at which this value is at maximum. This means that the filter orientation is fitted for line-like structures of both positive and negative contrast. The intensity $W_\sigma(\theta_{max})$ at the selected scale is stored for each pixel, to allow selection of pixels in a given range of intensities at a later stage of the processing.

2.3 Features for detection of stellate patterns

Two features have been defined for detection of stellate patterns of straight lines. These features are derived from the map of pixel orientations ϑ_i determined as described above. As these orientations reflect the line structure of the image, embedded in a background of random orientations, stellate patterns may be detected in this map by an increase of the number of pixels directed towards a central area.

The features are defined to quantify such an increase. At a given site i they are calculated from the orientations of pixels in a neighborhood N_i representing pixels j with distances $r_{ij} \in [r_{min}, r_{max}]$ from i. Not all pixels in this neighborhood are evaluated. Using a selection criterion on the intensity of the filter output $W_\sigma(\theta_{max})$, a subset of pixels S is determined representing potential sites of interest. Sites with negative contrast and sites where the intensity of the line feature is very low are excluded. To calculate the features, all pixels in the neighborhood of the test site i that belong to S and that are directed towards a disk of radius R centered at i are counted as a function of the direction φ in which they are located (Figure 2.2). Dividing the space around i into K bins of different directions, the number of neighboring pixels $n_{i,k}$ in bin k that are oriented towards the center can be calculated by

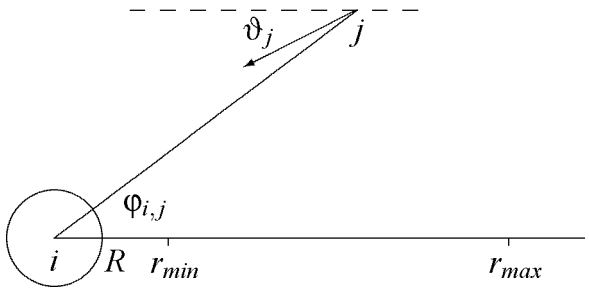

Figure 2.2: Pixels j that are located at a distance $r_{i,j}$ between r_{min} and r_{max} and that are directed towards a disk of radius R centered at the test site i are counted.

$$n_{i,k} = \sum_{j \in N_{i,k} \cap S} h(\vartheta_j, \varphi_{i,j}, r_{i,j}) \tag{2.6}$$

with the set $N_{i,k}$ denoting the neighboring pixels in direction φ_k, and

$$h(\vartheta_j, \varphi_{i,j}, r_{i,j}) = \begin{cases} 1 & \text{for } |\varphi_{i,j} - \vartheta_j| < \frac{R}{r_{i,j}} \\ 0 & \text{else} \end{cases} \tag{2.7}$$

The first feature which is defined is the total number of pixels with directions pointing to the center n_i, which can be calculated by summation of $n_{i,k}$ over k. In order to normalize this feature, the mean value of n_i and its variance are estimated under the assumption that the pixel orientation map is a uniformly distributed random noise pattern. The mean probability p that a pixel in this random map is pointing to the test site i is

$$p = \frac{2}{\pi N_i} \sum_{j \in N_i \cap S} \frac{R}{r_{ij}} \tag{2.8}$$

with N_i the total number of pixels in $N_i \cap S$. The normalized feature then is defined by

$$f_{1,i} = \frac{n_i - pN_i}{\sqrt{N_i\, p(1-p)}} \tag{2.9}$$

Because of the normalization, the sensitivity of the feature and its range do not change systematically when the neighborhood or target size R are changed. This enables changing these parameters adaptively and avoids problems at the edge of the image.

If an increase in the number of pixels oriented towards a region is found in a few directions only, it is not very likely that the site being evaluated belongs to the center of a stellate pattern. On the other hand, if evidence for spicules is found in many directions this should increase the likelihood of a stellate structure being present. To represent this property, a second operator is constructed. In each direction bin k, the mean probability of finding $n_{i,k}$ pixels oriented to the center out of a total of $N_{i,k}$ is calculated by applying (2.9) for each bin separately. Using binomial statistics it is determined how many times $n_{i,k}$ is larger than the

median value calculated for random orientations. Denoting this number by n_+ and with K' the number of bins, the second feature is defined by

$$f_{2,i} = \frac{n_+ - K'/2}{\sqrt{K'/4}} \tag{2.10}$$

where $K'/2$ is the expected value of n_+ when no signal is present. To avoid boundary effects, only bins in which a minimum number of contributing sites is present are considered. Therefore, near the breast edge the actual number of bins K' that contribute is smaller than K. The standard deviation of random fluctuations in the denominator normalizes the expression. It is noted that implementation of (10) is not as straightforward as it seems, as the median of a binomial distribution is not well defined. We perform a linear interpolation in between integer values of n to map the discrete distribution to a continuous function, which is regarded as a probability density function to determine the median m_k for each bin k. In case the distance between the median and the number of pixels directed to the center $n_{i,k}$ is smaller than 0.5, the increment used to calculate n_+ is taken as $n_{ik} - m_k + 0.5$, instead of the unit or zero increment used in all other cases.

To demonstrate the performance of the two operators an artificial stellate pattern was generated, which was hidden by adding i.i.d. Gaussian noise. The left column of Figure 2.3 shows an example, generated at a signal to noise ratio of 1.0, and the two corresponding feature images. Both operators yield a high output at the center of the test pattern. Calculation of the features was performed by applying the directional derivative filters at scales $\sigma = 1, 2$, and 4, where the width of the spicules was about 4 pixels. The test pattern in the right column of Figure 2.3 was generated at a signal to noise ratio of 0.25. Even in this very noisy background the feature f_1 still shows an increase at the position of the stellate signal. Also note that, in spite of a strong reduction of neighborhood size at the image boundary, there are no significant artifacts.

2.4 Application to mammograms

Stellate lesions in mammograms have a much more complex appearance than the artificial images of Figure 2.3. To deal with this, some modifications of the feature calculation were implemented and the mammograms were preprocessed by the following operations: First the breast image was shifted away from the matrix boundary, and an automatic procedure was applied to segment the breast area from the pectoral muscle in the oblique views. Next, the region around the breast tissue was replaced with a smoothed average of the neighboring breast tissue. In addition, in the region near the breast skin line the intensity fall off due to a decrease of the tissue thickness is corrected for, allowing embedding of the breast tissue in a more or less homogeneous background. By way of example, the upper row of Figure 2.4 shows a mammogram before and after preprocessing. For the breast edge correction a smoothed version of the image is calculated using a square uniform kernel of 3 mm in width. The correction is applied at all sites where this smoothed image I_s is smaller than a threshold T_e. At these sites the original pixel value I is replaced by $I' = I - I_s + T_e$. The threshold T_e is calculated as the mean pixel value in the inner part of the breast tissue region, determined by eroding the tissue area to 90% of its original size.

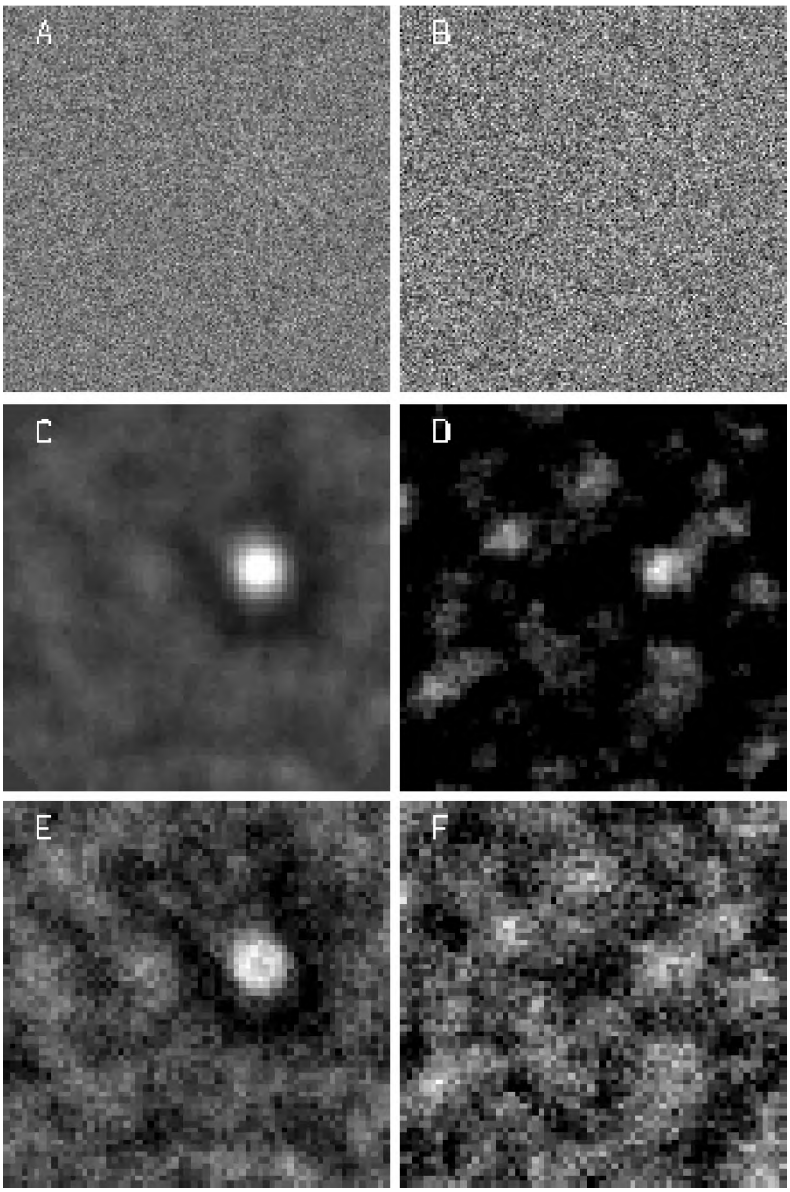

Figure 2.3: Figure (a) shows a test pattern embedded in Gaussian noise at a signal to noise ratio of 1.0. The features f_1 and f_2 constructed to detect the pattern are shown in Figures (b) and (d). Both features have a strong peak at the center of the stellate pattern. Figures (d),(e) and (f) show the same sequence for a pattern generated at a signal to noise ratio of 0.25. Still a peak is found in f_1 but the signal is lost in f_2.

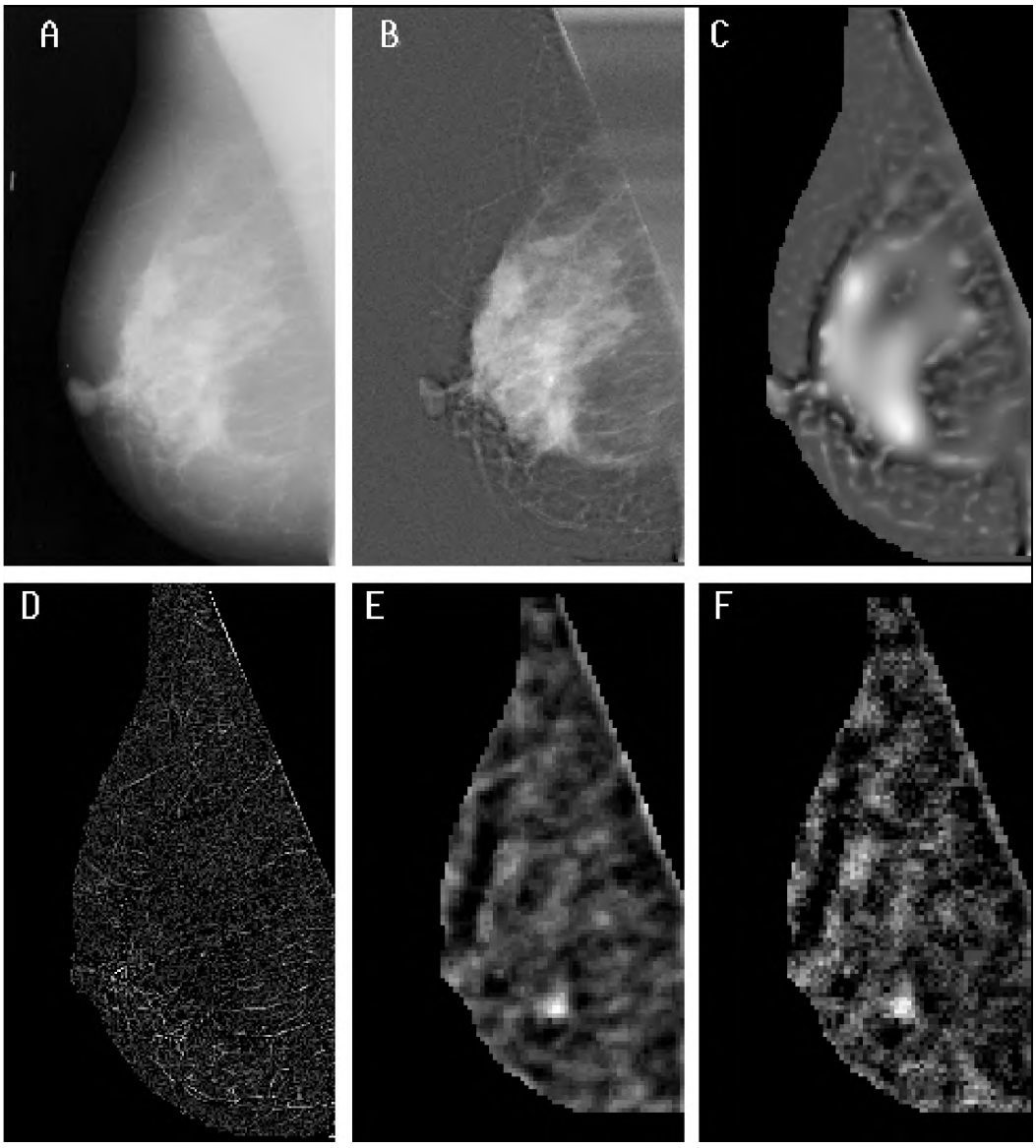

Figure 2.4: Image mdb120 of the MIAS database, showing a malign stellate distortion (a). In (b) the mammogram is shown after preprocessing, and (c) shows the output of the Laplace operator maximized over a range of spatial scales. In (d) the intensity of the line map is shown. The orientation map of all pixels with a positive output on the line intensity map is used to calculate the feature images f_1 and f_2 shown in Figures (e) and (f), where the Laplace output is used to adaptively set the size of the neighborhood. Both features have a strong peak at the position of the tumor. The cancer is detected even at a sensitivity level of 1 false positive in 50 images.

A modification of the feature calculation was introduced to deal with the fact that stellate patterns vary a lot in size. In mammograms, malignant stellate patterns often have a central mass. If this mass is small, the spiculated area normally is closer to the center than in case of a larger mass. Therefore, if a central mass is likely to be present at a given site, its estimated size can be used to set the size of the neighborhood to be evaluated for spicules. This idea was used by testing for the possible presence of a mass by application the the Laplace operator ∇^2_σ at a number of different spatial scales. If the output of this operator exceeded a certain threshold T_L the neighborhood size parameter r_{max} was scaled with σ_{max}, the scale level with the highest output. To avoid artifacts near the matrix boundary, the Laplace output at sites i was taken to be invalid if the distance between i and the matrix boundary was less than 2σ. In Figure 2.4 the maximum of the Laplace operator over a number of scales ranging from σ = 3 to 8 mm is shown. It is remarked that the output of the Laplace operator was greatly improved by the preprocessing steps described in the first part of this section.

It appeared that the presence of distinct, clear structures in the mammograms like blood vessels, skin folds or sharp outlines in the glandular tissue deteriorated the performance of the detection algorithm. This was caused by the fact that at strong boundaries between regions the second-order directional derivatives give ripples in parallel to the boundary. This leads to structure in the orientation map which is not related to the presence of lines in the range of chosen widths. To reduce this effect, at a coarse scale regions with a high gradient magnitude were determined, and within these regions pixels with orientations ϑ_i perpendicular to the gradient orientation (thus parallel to the boundary) were removed from the subset S of contributing sites. Gradient magnitude and orientation were calculated by using first order Gaussian derivatives at a relatively large value of σ for differentiation.

2.5 Combining features for classification

A number of different classification methods were applied to combine the two features f_1 and f_2 into a single measure of suspiciousness. This measure can be used to generate prompts to alert radiologists. The classifiers were built using a data set of example feature vectors taken from 14 digitized mammograms, all showing a stellate lesion labeled by an expert radiologist. Only pixels inside the central region of a lesion were labeled as abnormal, not the spicule pattern itself. With an erosion procedure, these labeled regions were all reduced to the same size, to give each image an equal weight in the training procedure. The training set was constructed using all malignant-labeled pixels and a large amount of randomly selected normal background points. Several non-parametric classification techniques were used: Bayesian with non-parametric estimation of the probability densities, k-Nearest-Neighbors (kNN), a decision tree and a neural network [10]. Each of these was implemented in such a way that the output at each pixel could be interpreted as a measure of suspiciousness. The classifiers are described in more detail in the rest of this section. Figure 2.5 gives an impression of the training data in feature space. It shows the estimated probability density functions for pixels in both normal and abnormal areas.

The Bayesian technique is based on estimation of the class conditional probability density functions of the two features for background and tumor points. At a chosen level of the prior probabilities for both classes, the two features can be combined using Bayesian decision theory, giving each pixel a likelihood of malignancy. A non-parametric method for

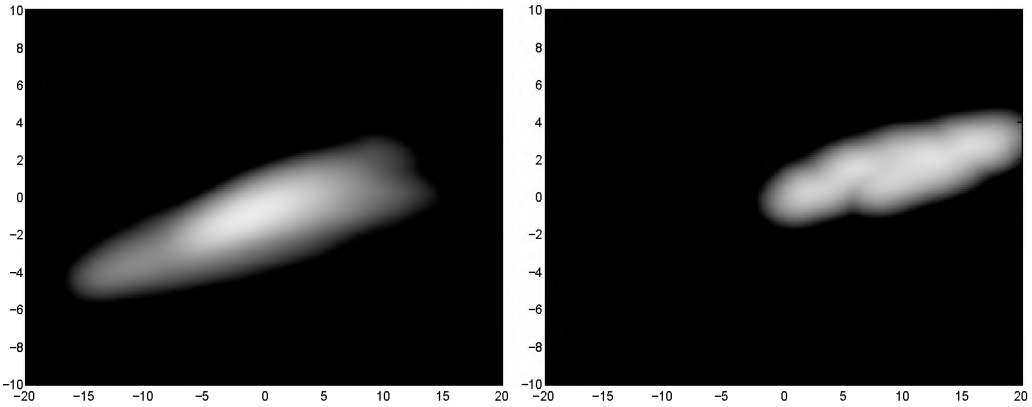

Figure 2.5: Estimated probability density functions (pdf) calculated from a set of 14 mammograms, each showing a stellate malignancy. The pdf for pixels in normal tissue is shown in the upper figure, the lower shows the pdf for pixels in the central part of a malignant stellate pattern. Feature f_1 is plotted horizontally and f_2 vertically.

density estimation was used.

To obtain a measure of suspiciousness using kNN, the number of malignant vectors among the k most 'similar' vectors in the training set was used. It is noted that this number depends on the a priori chances of the malignant and benign classes, which could be manipulated varying the number of patterns of both classes in the data set. Because this method is very slow, a much smaller training set was used than with the other methods.

A decision tree recursively subdivides regions in feature space into two subspaces, each time using a threshold in one dimension that separates the two classes 'as much as possible'. For a given subspace the process stops when it only contains patterns of one class. In our implementation we used the ID3 information criterion [10] to determine threshold values from the training data. As an additional stopping criterion, we used a threshold on the number of points a subspace, to prevent the method from overfitting. At each end node of the tree, the probabilities p_m for both malignant and p_b for benign were estimated by calculating the number of training feature vectors of both classes in the subspace corresponding to the node, and by respectively weighting these numbers by the number of malign and benign training patterns. To classify a pixel, the tree is used to determine the end node or subspace corresponding to its feature vector. The likelihood ratio p_m/p_b at this node is used as a measure of suspiciousness.

Another classification technique that was used is a feed-forward neural network with one hidden layer, trained with the back-propagation algorithm. The number of hidden nodes was 5, but hardly any variation in performance was noticeable when we changed this number. Each benign feature vector was labeled 0, each malignant vector 1. After training, this resulted in a network that gave values close to one for suspicious vectors, and values close to zero for benign vectors.

2.6 Experimental set-up and performance measurement

The methods were applied to mammograms taken from the MIAS database [13]. All malignant stellate lesions (9) and architectural distortions (10) were selected. By adding the first

31 normals from the database (003 to 046) a test set of 50 images was obtained. Calculation of the features was performed on 1k×1k images. Images from the MIAS database were reduced to this size by sub-sampling the original 50 micron/pixel data, merging 4×4 blocks of 8 bits pixels into one 12 bits pixel by addition. Subsequently, adaptive noise equalization [6] was applied to obtain a uniform noise level over the image. After this transform, the image data was reduced back to 8 bits/pixel. By using noise equalization, the requantization error introduced by this reduction could be kept small. Calculation of stellate features was performed for sites inside the tissue area sampled at 1.6 mm intervals. The two feature images f_1 and f_2 were slightly smoothed before combining them in a classifier, to reduce the influence of noise on the subsequent process of marking suspicious areas.

The pixel classifiers were constructed using the training set of 14 mammograms, which were not used for testing the methods. To measure the performance of a classifier, the likelihood images it produces were converted to binary images by using a threshold T_l to mark the most suspicious pixels. From these marked pixels regions were formed by applying a morphological closing and opening using a 3×3 structuring element. Subsequently, regions smaller than 500 pixels were removed. This minimum of 500 pixels corresponds to the size of a circular area of 0.5 cm in diameter. The remaining regions were compared with the annotated true masks to determine detection performance. For this purpose, within each region the position of the maximum of the likelihood image was determined. If this maximum was inside the annotated true lesion, the lesion was regarded as detected, otherwise a false positive was counted. This criterion was chosen because the pixel labeled with maximum probability of suspiciousness in a marked region can well be used as the location to generate a prompt for the human observer, and then the only thing that matters is whether or not this prompt is at the right position. It is noted that this criterion is different from the one used earlier in [5] based on the overlap of the deleted and annotated regions, to avoid problems in extreme cases when very large regions were marked as detected. Variation of the threshold T_l level used to mark suspicious regions allows generation of prompts at different sensitivities. Results obtained in that way are presented as FROC curves in which the true positive fraction is plotted as a function of the average number of false positives per image.

Parameter settings were chosen by optimizing the output of the algorithm on the training set. For estimation of line orientation the directional second order derivatives were applied at three spatial scales $\sigma = 0.1$, 0.17, 0.29 mm. These scales cover the range of widths of the spicules as occurring in our datasets. The gradient operator used to remove ripples at strong gradients was computed at $\sigma = 1$ mm. Pixels for which the gradient magnitude was larger than $T_g = 25$ and for which the difference between line and gradient orientation was smaller than $\pi/6$ were not considered for calculation of f_1 and f_2. To determine the neighborhood size the Laplace operator was applied at σ = 3.2, 4.0, 5.4, 6.2, and 7.8 mm. At these scales the mammographic tumor masses that we are interested in are represented. Fixed values $r_{min} = 4$ mm and $r_{max} = 16$ mm were used when the Laplace operator had a low output. For sites at which the output of the Laplace operator was higher than $T_L = 20$ the parameter r_{max} was set to $3\sigma_{max}$. For counting the number of pixels oriented to the center $R = 2$ mm was used. The number of direction bins K used to compute f_2 was set to 24. It appeared that the algorithm was not sensitive for changes of K in the range of 20 to 30. For high values of K the number of pixels in each bin gets to small to use binomial statistics in a valid way. For the image presented in Figure 2.4 the intensity of the line-map and the two features are shown as an example. A fairly obvious case was selected to guarantee visibility of the tumor

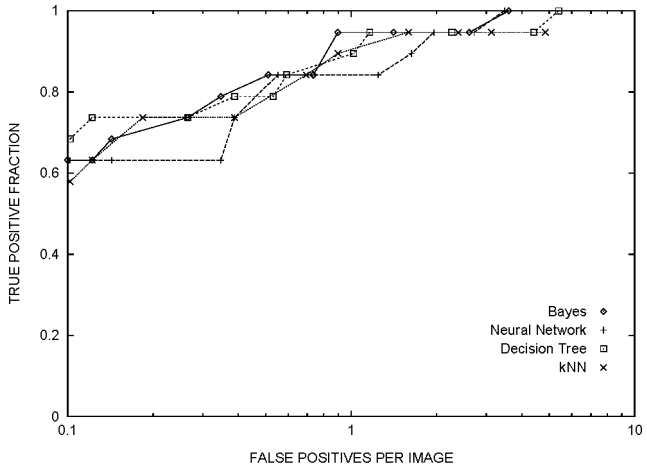

Figure 2.6: FROC curves for detection of stellate abnormalities using different classifiers. Curves are computed for 50 images of the MIAS database, including 19 malignant stellate patterns and 31 normals.

in the reproduction. The cancer was not missed by any of the classification methods even at a specificity of 0.02 false positive/image.

Results obtained on the MIAS database are shown in the FROC curves in Figure 2.6. For each classifier a curve is shown. For kNN, the decision tree and the neural network, the curves represent the average performance of a number of different classifiers, obtained by training with a different selection of background points. For the Bayesian classifier all background points were used for training. It appears that differences between the classification methods are small.

Figure 2.7 shows a few of examples, all representing an area of 5 cm in diameter. The spiculated mass in A was detected at all specificity levels up to 0.02 FP/image. The mass in B could only be detected at 4 FP/image. The two architectural distortions in C and D were detected at respectively 0.1 and 0.25 FP/image and higher. In E and F two false positives are shown that were still picked up at 0.1 FP/image.

To study the effect of using a multi-scale method for estimating line orientation FROC curves were calculated for single scale cases as well, at each of the three scales used. Results can be compared to the multi-scale case in Figure 2.8. Also the importance of using adaptive neighborhood sizes, based on the Laplace operator output, and the restricted use of pixels at strong gradients was investigated. Figure 2.9 shows the FROC curve obtained without using these two optimizations. It was found that on average about 50% of the tissue pixels contributed to calculation of the features. For these pixels the intensity of the line operator exceeded a small threshold used to select lines with positive contrast. Experiments with an increased value of the threshold to ignore pixels at which the line operator output was low did not improve the results.

2.7 Discussion

The results show that on the basis of a line-based pixel orientation map many subtle spiculated lesions and architectural distortions can be detected at a high degree of specificity. In

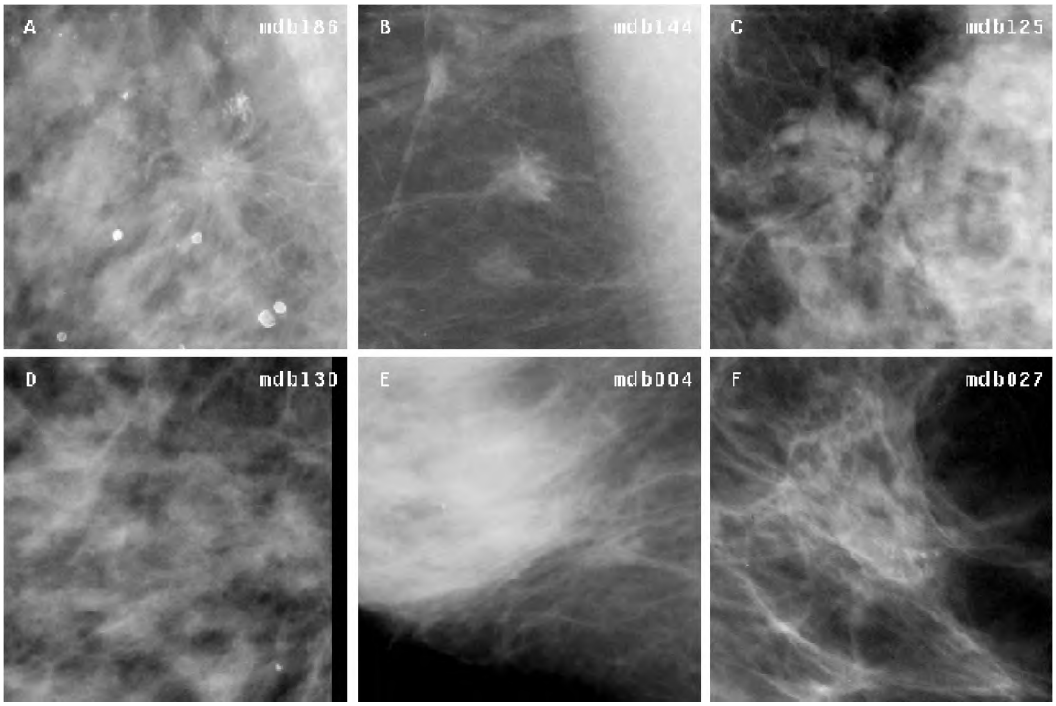

Figure 2.7: Examples taken from the MIAS database, representing patches of 5×5 cm. The spiculated mass in (a) was detected at all specificity levels up to 0.02 FP/image. The mass in (b) could only be detected at 4 FP/image because only very short spicules are present. The two architectural distortions in (c) and (d) were detected at respectively 0.1 and 0.25 FP/image and higher. In (d) the center of the tumor is on the central row at the right edge. Only vague thin lines are visible. In (e) and (f) two false positives are shown that were still prompted at 0.04 FP/image.

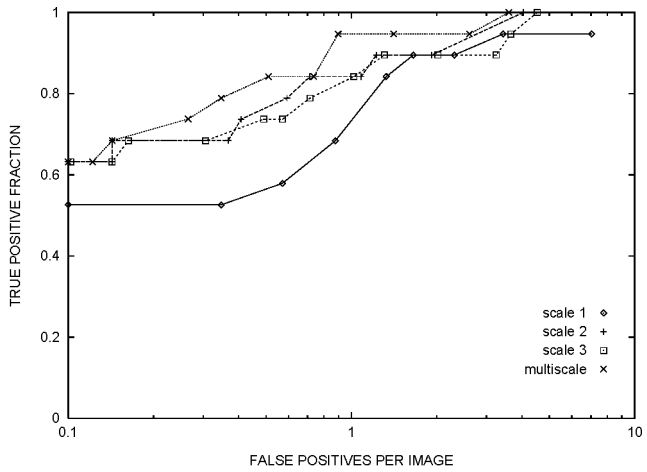

Figure 2.8: FROC curves obtained with line orientations calculated at different spatial scales $\sigma = 0.1, 0.17, 0.29$ mm, and with a multi-scale approach combining the three scales. Curves are computed for 50 images of the MIAS database, including 19 malignant patterns. The Bayesian classifier was used.

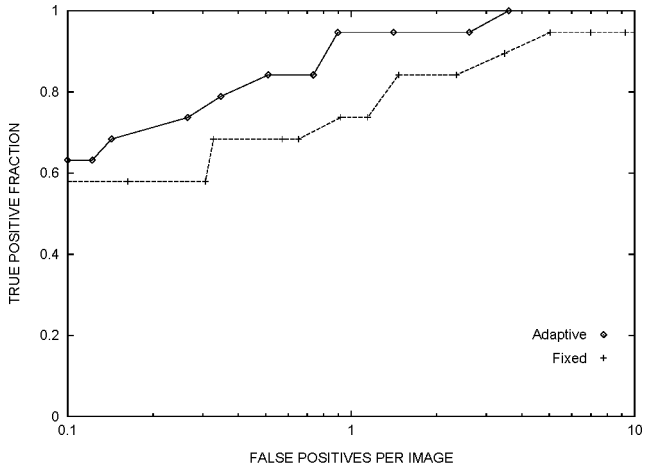

Figure 2.9: FROC curves showing the difference between adaptive and fixed parameter settings. In the adaptive mode the neighborhood size evaluated for spicules is selected in proportion to the size of a central bright area, if such an area is likely to be present, and pixels that may introduce artifacts due to the presence of a strong step edge are excluded if the line orientation estimator yields a direction in parallel to the edge.

the test set, spiculated lesions are more easily detected than architectural distortions. Lesions which are lost first when the sensitivity is reduced are those without many spicules or those in which spicules are very short. For instance, the lesion in Figure 2.7b (mdb144), which has the appearance of a small irregular mass, is missed first. To recognize such lesions, features for signaling the presence of a central mass must be added.

The computations are based on a new method for estimation of line orientation described in section 2. Given the fact that many of the stellate patterns in the test set are quite faint and noisy the method appears to be very robust. Attempts to improve results by using a threshold on the line operator output were not successful. This indicates that even at sites where the presence of a line seems rather unlikely the estimated orientations still have a positive contribution on average. Figure 2.8 shows that using the multi-scale approach yields better results than using one spatial scale only. The performance reduction is limited when using line orientations estimated at $\sigma = 0.17$ mm and $\sigma = 0.29$ mm, but there is a strong a strong reduction when using only the highest spatial resolution ($\sigma = 0.1$ mm). It may the case that spicules at this high resolution occur less frequent. It should also be considered, however, that the image resolution we have used (0.2 mm/pixel) was too low for accurate calculation of the Gaussian derivatives at this high resolution.

Figure 2.9 shows that the use of an adaptive neighborhood based on the size of a central mass, if present, and the use of rule to avoid artifacts at strong edges improves performance importantly. The number of false positives per image at a given level of sensitivity is roughly reduced by a factor of three by applying these optimizations. Results obtained by using different classifiers were similar. However, one should note that decision trees and neural networks are much more suited to scale up to higher dimensional data than k-nearest neighbor and the non parametric Bayes classifier. In future research this will become important, because features will be added to improve performance. In fact, it is not very surprising that there is not much difference between the classifiers when one looks at Figure 2.5. The

class conditional probability density functions are compact and only have limited overlap. In such cases classification is relatively easy and most techniques should work well. Inspecting the pdf of the background pixels in Figure 2.5 more closely reveals that the maximum of this function is not at the origin. On average the values of f_2 are smaller than zero. This is caused by the fact that normalization is performed for the reference condition that neighborhood pixels have random orientations. Of course, this is not true for the line-based pixel orientations estimated with the method described here. Nearby pixels will be correlated due to the scale of the differential operators that were used, and due to the structure of mammographic patterns. Because of this correlation, also the variances of the probability density functions are larger than one. For the purpose of normalization, however, the use of uncorrelated random orientations as a reference turned out to work well.

Calculations of the Gaussian derivatives are implemented in the Fourier domain, which turned out to be rather time consuming. Processing time on a HP712 workstation was about 20 minutes/image, where 80% of the processing time was used for Fourier transforms. The use of the AFT algorithm also restricts application to image dimensions which are a power of two. By approximating the kernels in the spatial domain, using a pyramid representation of the data, the algorithm can be speeded up. This will also enable the use of a higher resolution at unrestricted image dimensions, without requiring an excessive amount of memory.

Bibliography

[1] G D Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Anal.*, 36:1169–1179, 1988.

[2] L M J Florack, B M ter Haar Romeny, J J Koenderink, and M A Viergever. Scale and the differential structure of images. *Image and Vision Computing*, 1992.

[3] A K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[4] N Karssemeijer. Recognition of stellate lesions in digital mammograms. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 211–220. Elsevier, Amsterdam, 1994.

[5] N Karssemeijer. Detection of stellate distortions in mammograms using scale space operators. In Y Bizais, C Barrilot, and R Di Paola, editors, *Information Processing in Medical Imaging*, pages 335–346. Kluwer, Dordrecht, 1995.

[6] N Karssemeijer, J T Frieling, and J H Hendriks. Spatial resolution in digital mammography. *Invest radiol*, 28(5):413–9, May 1993.

[7] W P Kegelmeyer. Computer detection of stellate lesions in mammograms. *SPIE 1660*, pages 446–454, 1992.

[8] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[9] J J Koenderink and A J van Doorn. Generic neighborhood operations. *IEEE PAMI*, 1991.

[10] D Michie, D J Spiegelhalter, and C C Taylor. *Machine learning, neural and statistical classification*. Ellis Horwood, 1994.

[11] S L Ng and W F Bischof. Automated detection and classification of breast tumors. *Comput Biomed Res*, 25:218–237, 1992.

[12] C J Savage, A G Gale, E F Pawley, and A R M Wilson. To err is human; to compute divine? In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 405–414. Elsevier, Amsterdam, 1994.

[13] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[14] J A M van Dijck, L M Verbeek, Hendriks J H C L, and R Holland. The current detectability of breast cancer in a mammographic screening program. *Cancer*, 72:1933–1938, 1993.

Chapter 3

Features for mass detection

3.1 Introduction

In Chapter 2, a method was described to detect stellate lesions and architectural distortions based on the detection of the radiating pattern of spicules [1]. Although approximately half of the lesions in the screening show spiculation, many are only detectable by the mass. The spicules were detected using a statistical analysis of line orientations. A similar approach was used for detection of masses: statistical analysis of gradient orientations. To detect the whole spectrum of masses, both spiculation features and mass features are required.

An experiment was done to examine the contribution of both type of features to the detection of a consecutive set masses from the Nijmegen screening program. Three neural networks were trained: the first using spiculation features, the second using mass features and the third using both type of features. The next section describes the features that were developed for mass detection. In Section 3.3 the experiment is described.

3.2 Features for detection of masses

For detection of masses a similar approach is taken as for detection of spicules. Instead of the map of line-orientation, we now use a map of gradient orientations. Pixels that are inside a mass will be surrounded by pixels with a gradient orientation towards the central pixel. If no structure is present, a random direction is found. Statistical analysis of this map makes it possible to find masses. Two features similar to the features for spicule detection are developed.

By convolving the image with two first derivatives of the Gaussian, we compute I_x and I_y, the gradients in the x and y directions. A Gaussian with a sigma of 1 mm was used in this experiment. With the formula

$$\theta(x,y) = tan^{-1}\left(\frac{I_y}{I_x}\right) \tag{3.1}$$

we can compute the direction of the gradient, and with

$$I = \sqrt{I_y I_y + I_x I_x} \tag{3.2}$$

the magnitude.

Two new features can now be defined in the same way as for spicule detection. The first represents the number of pixels with an intensity gradient pointing towards the central pixel, the second feature whether these points occur in all directions of the central pixel. At a given site i these 2 features are calculated from the orientations of pixels in a circular neighborhood of pixels with a distances between r_{min} and r_{max} of i. The gradient orientation operator gives a magnitude value, and all points with a gradient larger than a (low) threshold are included in the computation.

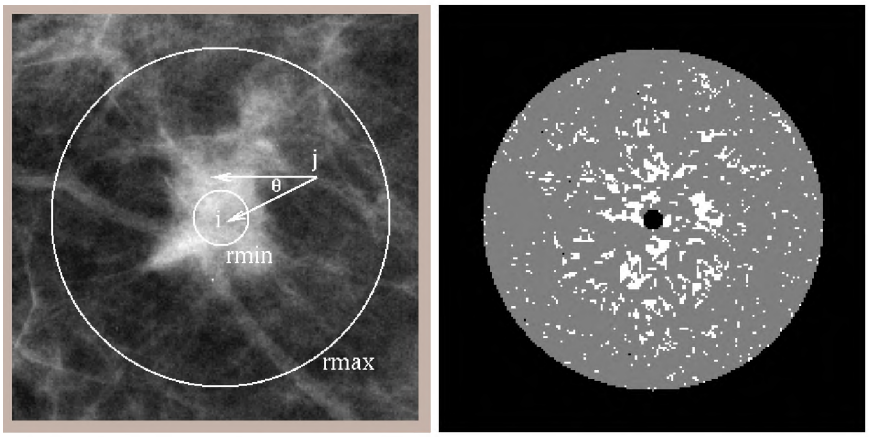

Figure 3.1: Pixels j located at a distance r_{ij} between r_{min} and r_{max} and directed towards the test site i are counted. Left, a mass is shown, in the right figure the white pixels are the pixels with a gradient orientation towards the center i, the grey pixels are inside the neighborhood region but are not oriented towards the center.

In Figure 3.1 the method is explained. A measure of suspiciousness for the pixel in the center i is computed. For all pixels j with at a distance between r_{min} and r_{max} of i, it is determined whether the gradient orientation vector is oriented towards the center. A pixel is considered to be oriented towards the center if the angle θ between the line through i and j and the gradient orientation is smaller than a constant D divided by the distance between i and j. A small D will reduce the number of neighbor pixels that is oriented towards the center. In Figure 3.1b all grey and white pixels are in the circular neighborhood, where the white pixels are oriented towards the center.

If no structure is present, the probability p_{ij} that a pixel j is randomly pointing towards the center i is given by

$$p_{ij} = \frac{D}{d_{ij}\pi},$$

where d_{ij} is the distance between i and j. The mean probability that a pixel in the neighborhood is oriented towards the center i is

$$p_i = \frac{1}{N}\sum_j \frac{D}{d_{ij}\pi},$$

with N_i the number of pixels in the circular neighborhood. Therefore, for each circular neighborhood we can compute the expected number of points with an orientation towards the center $N_i p_i$, and its variance $\sqrt{N_i p_i (1-p_i)}$. The number of pixels n_i that is actually

pointing towards the center is counted, and the feature representing suspiciousness is computed by

$$g_{1,i} = \frac{n_i - N_i p_i}{\sqrt{N_i p_i (1 - p_i)}}.$$

This feature is adaptive to varying neighborhood sizes and D, and does not suffer from artifacts near the edge of the breast, where the circular neighborhood lies partly outside the breast area.
If an increase of the number of pixels oriented towards a region is found in a few directions only, it is not very likely that the site being evaluated belongs to a mass. On the other hand, if gradient directions away from the center are found in all directions, this should increase the likelihood of a mass being present. To represent this property, another feature is constructed.

The space around i is divided like a pie into K bins. The number of neighboring pixels in bin k that are oriented towards the center pixel i is denoted by $n_{i,k}$. For each bin, the number of pixels in this bin $N_{i,k}$ and the mean probability p_i that a random pixel is directed towards the center are known. To avoid boundary effects, only bins containing more than a specified minimum number of contributing sites are considered, giving K' bins. The number of pixels that is directed towards the center is binomially distributed with $B(N_{i,k}, p_i)$. The median value of this distribution is computed, and it is determined how for how many bins $n_{i,k}$ is larger than the median. Denoting this number by n_+ and with K' the number of used bins, the feature is defined by

$$g_{2,i} = \frac{n_+ - K'/2}{\sqrt{K'/4}} \tag{3.3}$$

where $K'/2$ is the expected value of n_+ when no signal is present. The standard deviation of random fluctuations in the denominator normalizes the expression. If gradients are found in many bins, n_+ will be a high number and this feature will indicate a circular density.

In Figure 3.2 a mammogram with a mass is shown. The features f_1 and f_2 are described in Chapter 2. The mass gives a small signal for the spiculation features and a large signal for the mass features.

3.3 Experiment

Three different neural network were trained using 39 images form the MIAS dataset [2]. The first neural network was trained using spiculation features, the second neural network using mass features, the third neural network was trained using both spiculation and mass features. Simple 3-layer feed-forward neural networks with 5 hidden nodes were trained for this purpose. To networks were trained using 39 images from the MIAS Dataset [2], one of the first publicly available data sets.

A data set with 132 mammograms containing a malignant abnormality, 132 normal contra-lateral mammograms and 208 normal mammograms was used to test the algorithms. The 132 masses form a consecutive set of all cancers that were detected in two years in the Nijmegen screening program, only cases were microcalcifications were the only sign of malignancy were excluded. The results are presented in FROC curves: on the horizontal axis

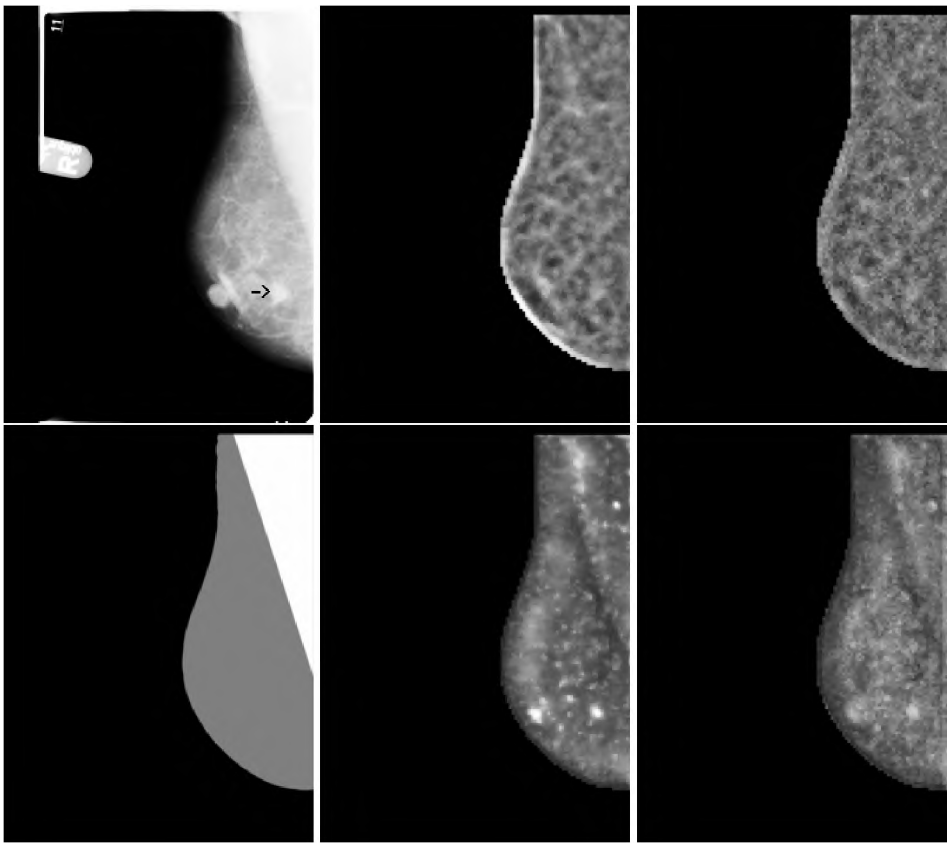

Figure 3.2: Mammogram with a mass indicated by the arrow. Top line: the mammogram, and the features f_1 and f_2, bottom line the segmentation of the breast and the features g_1 and g_2. Using the 2 spiculation features this mass is detected at a specificity level of 4.4 false positives per image, using the mass features at 0.4 false positives per image, and at 0.04 false positives using all features. The skin line causes an artifact in the f_1 feature image. Note the signals in the g_1 and g_2 images at the nipple and for a visible lymph node.

the average number of false positives per image is given, on the vertical axis the percentage of tumors that is detected.

In Figure 3.3, the FROC curves for the three neural networks are shown. On this data set, the neural network that classifies using the spiculation features performs rather poor, because most masses in this data set do not have clear spiculation. The neural network that is trained using the mass features performs much better, but combining both mass characteristics is required for optimal results. Over 80% of the masses are detected at 1 false positive per image.

3.4 Conclusions

Combining both mass and spiculation features yields the best classifier for a dataset with a variety of mass types. The relatively low curve for the spiculation features shows that most masses in the set are vague or circumscribed masses without spiculation. Therefore, the curve for the neural network that was trained using spiculation features is much lower than reported in the experiment in Chapter 2, where all cases in the test set were spiculated.

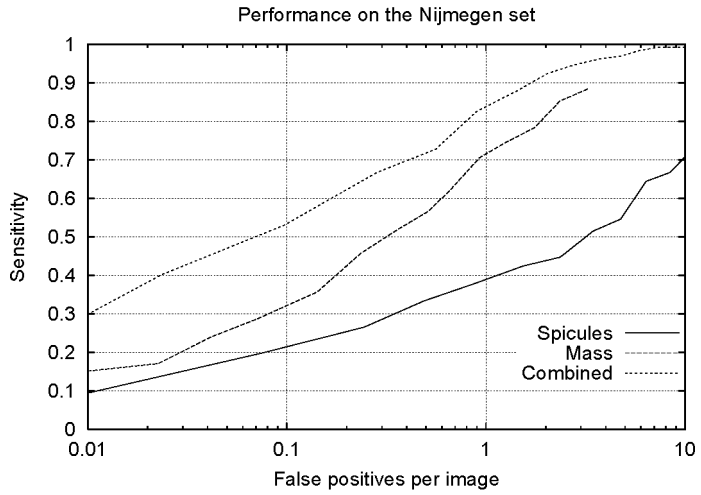

Figure 3.3: The performance of the three neural networks. The combined version clearly outperforms the mass or spiculation network.

To detect the whole spectrum of masses and distortions, both spiculation features and mass features are required to achieve high sensitivity at reasonable specificity levels.

Bibliography

[1] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[2] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[3] G M te Brake and N Karssemeijer. Detection of stellate breast abnormalities. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 341–346. Elsevier, Amsterdam, 1996.

Chapter 4

Single and multi-scale detection of masses in digital mammograms[1]

Abstract

Scale is an important issue in the automated detection of masses in mammograms, due to the range of possible sizes masses can have. In this work, it was examined if detection of masses can be done at a single scale, or whether it is more appropriate to use the output of the detection method at different scales in a multi-scale scheme. Three different pixel-based mass detection methods were used for this purpose. The first method is based on convolution of a mammogram with the Laplacian of a Gaussian, the second method is based on correlation with a model of a mass, the third is a new approach based on statistical analysis of gradient orientation maps.

Experiments with simulated masses indicated that little can be gained by applying the methods at a number of scales. These results were confirmed by experiments on set of 71 cases (132 mammograms) containing a malignant tumor. The performance of each method in a multi-scale scheme was similar to the performance at the optimal single scale. A slight improvement was found for the correlation method when the output of different scales was combined. This was especially evident at low specificity levels.

The correlation method and the gradient orientation analysis method have similar performance. A sensitivity of approximately 75% is reached at a level of 1 false positive per image. The method based on convolution with the Laplacian of the Gaussian performed considerably worse, in both a single and multi-scale scheme.

4.1 Introduction

It is well known that screening mammography is a difficult task for radiologists and that screening errors are hard to avoid. Retrospective studies show that in current breast cancer screening between 10% and 25% of the tumors are missed by the radiologists [22, 2, 3, 7]. One of the signs that has to be detected in mammograms are masses. Masses can be hard to detect because, due to the projection, they are often partially covered by glandular tissue.

[1]Published as: G.M te Brake, N. Karssemeijer, *Single and multiscale detection of masses in digital mammograms*, IEEE Transactions on Medical Imaging, vol 18, nr. 7, 1999.

Recent work has shown that many of the tumors that are missed by radiologists can be detected by a system that automatically detects masses [21]. A Computer Aided Diagnosis (CAD) system that prompts suspicious regions can draw the attention of the radiologist to a tumor he might otherwise overlook [8, 5, 1, 12].

One of the problems in automated detection of masses is the choice of the scale that should be used. Masses vary largely in size, ranging from a few millimeters to a few centimeters. Only a few publications on automated detection of mammographic masses address the issue of scale, and most of these have been validated on a very small dataset. Brzakovic and Neskovic [4] applied their algorithm which is based on fuzzy pyramid linking on a number of scales, to detect abnormal structures over a range of sizes. Ng and Bischof [17] detect the central mass of lesions using a basic template matching scheme which is applied on a number of scales. A circular Hough transform was used by Groshong and Kegelmeyer [6] to detect circumscribed masses. This transform looks for circular blobs with a radius between 3 mm and 30 mm, in a multi-scale approach. Their algorithm was tuned in such a way that small and large tumors will give a similar signal, and was tested on 22 mammograms containing circumscribed masses that are present in the MIAS dataset [20]. However, only four of them were malignant. Benign masses normally have a sharper boundary than malignant masses, which may be favorable for their approach because the orientations of the gradients at the edge of these masses can be determined more accurately. Nishikawa et al. [18] report a strong correlation between performance of their algorithm and the tumor size. Of all tumors smaller than 15 mm, only 30% were detected at 1 false positive per image, while of all tumors larger than 20 mm in size, 85% were found at 1 FP per image. However, Miller and Ramsey [16], using a multi-scale approach, report that the performance of their system does not depend on the size of the tumor. On a test set of screening-detected tumors and for all sizes approximately 60% of the tumors were detected at a specificity level where 25% of the women are falsely prompted (a false positive in at least one of the two views). An approach called directional recursive median filtering was used by Zwiggelaar et al. [25] to detect the central mass in stellate lesions. A one dimensional filter is applied to each pixel in a number of different orientations and on a number of different scales. Unfortunately, the benefit of detecting on multiple scales is not reported. In contrast to Miller and Ramsey, they report better results on large tumors than on small ones. Wavelet analysis is a promising approach to perform multi-resolution analysis of images. Wavelets were used by Li et al. [15] to incorporate multi-scale information in a mass detection method, but this work focused on directional information to detect spicules. A tree-structured wavelet transform was used to segment mass-like regions, but the gain of using multi-scale information instead of single-scale was not investigated. Most other wavelet papers in mammography focus on microcalcifications.

The aim of this paper is to examine the importance of scale for detecting the tumor mass in a more comprehensive way. Single and multi-scale methods will be compared on a large dataset. In a single-scale approach, the chosen scale should not be too small to avoid large masses to be missed, and not too large, to ensure high sensitivity for small tumors. The requirement for accurate setting of the scale parameter was examined for three detection methods. The possible advantage of detecting masses using the output of a detection method at multiple scales was also examined. It is remarked that multi-scale methods could also be applied to study properties of the tumor boundary or texture. Such methods were not investigated in this study.

Three methods for mass detection were used in this study, two standard methods from the field of image analysis and one novel method for mass detection. The first standard method is based on convolution with the second derivative of the Gaussian function, a Mexican hat shaped function. The second standard method is based on template matching. Using a model of the shape of a typical mass, the correlation between local image intensity values and the model is computed. Both methods have been used previously for detection of masses in mammograms by several other groups [17, 14, 9, 13, 19, 24]. The third method is an adaptation of an algorithm that was developed for detection of radiating patterns of spicules [10]. A gradient orientation map is computed and analyzed statistically to detect bright regions. These three methods will be described in more detail in the next section.

Multi-scale detection was performed in two ways. For each pixel in the image, the detection method was computed on a number of scales. The first approach combined the output over these scales using a neural network. The second method assigned each pixel the maximum output of the detection method over the range of scales. The performance of the methods in the multi-scale approaches will be compared to the performance of the methods on a single scale.

In Section 4.3 results of the methods on simulated masses are described. Simulated masses were projected on a mammographic background and used to examine the importance of choosing optimal values for the scale parameters for all three detection methods. The response of the methods for altering the size or intensity of the masses was examined quantitatively to gain insight into the robustness of the methods.

Masses were simulated according to the size distribution derived from a set of over 300 mass annotations in our database, in order to find a good value for the scale parameter.

In Section 4.4, results on real mammograms will be described. The detection methods were applied to a database of 71 tumors (132 mammograms), a consecutive set of all masses that were detected in two screening rounds in Nijmegen, the Netherlands. The effect of varying scale on the performance of the detection methods was examined on these real masses, and the three methods were compared.

4.2 Methods for mass detection

In this section, the methods for mass detection that were used in this study will be described. The first two methods are standard techniques in the field of image processing and analysis, the third method is a new approach. The three mass detection methods are pixel-based, which means that local image properties are computed at each pixel.

A mammogram is first preprocessed automatically to segment the breast from the background, and to compensate for the breast fall-off due to thickness variation [10]. Next, the desired features can be computed, giving a feature vector for each pixel. When a feature vector has been computed, it can be mapped to a measure of suspiciousness. If only one feature is used, its value can be used as the measure of suspiciousness, otherwise a classifier is required to compute a combined measure. When all pixels are given a measure of suspiciousness, a threshold is applied to segment suspicious regions. These regions can be used as the start for a second step where more regional features can be computed, like fuzziness of the boundary of the segmented area to remove false positive signals, but this was not part of this work.

4.2.1 Laplacian filtering

A common way to detect bright blobs in an image is by convolution with a zero-mean function which has a positive center and a negative surround. Several authors have used such a filter for detection of masses in mammograms [9, 13, 19, 24]. A common filter for this purpose is the second derivative of the Gaussian, called the Laplacian of the Gaussian (LoG). The filtered image is computed by

$$\text{Filtered image}(x,y) = \sum_{n=-N}^{N} \sum_{m=-N}^{N} I(x+m,y+n)LoG(m,n),$$

where N determines the size of the region in which the convolution is computed. Because of computational reasons, it was implemented as a Difference of Gaussians filter (DoG-filter)[23], where convolutions of two Gaussians with a different scale are subtracted. This difference measure is very similar in shape to the Laplacian of the Gaussian function if the ratio between the two sigmas of approximately 1.6. The equation shows that the output of the Laplacian filter scales linearly to the brightness of the mass. Because masses in mammograms differ largely in intensity this can be a drawback, because it makes faint masses hard to detect. An example of a mammogram and its feature image for the Laplacian convolution filter can be found in Figure 4.1a and 4.1b.

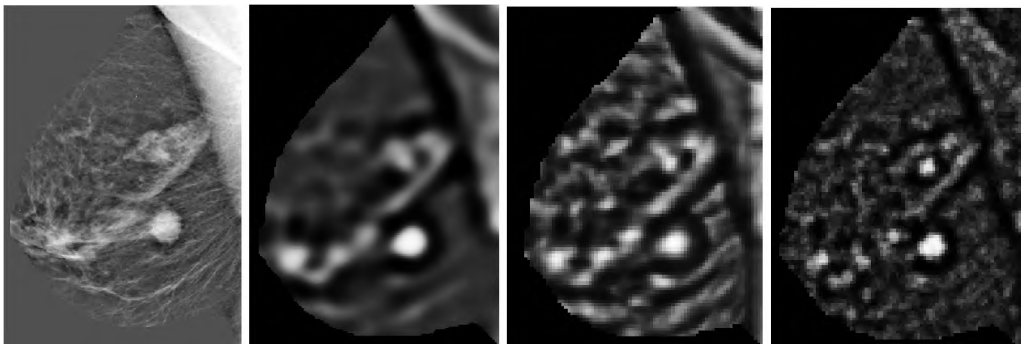

Figure 4.1: Mammogram with an obvious mass. For each of the three mass detection methods, its feature image at a medium scale is shown. (a) The preprocessed mammogram. (b) Feature image for the Laplacian convolution method. (c) Feature image for the template matching method (d) Feature image for the gradient orientation method.

The scale parameter for Laplacian filtering is the parameter sigma of the Gaussian. If the value of sigma is chosen appropriately, masses will give a higher output for this filter than normal tissue.

4.2.2 Template matching

If a model of a mass is assumed, a mammogram can be searched for regions resembling this model. This can be done by shifting a window across the mammogram, while locally computing the correlation measure between the overlapping region and the assumed model. This technique is known as template matching, and has been used in some of the earlier papers on detection of masses in mammograms [17, 14], where a rather simple model was used.

In our implementation, a mass with radius R is modeled in a circular template window with a radius of $1.3R$ using the template model function

$$T(x,y) = \begin{cases} R^2 - x^2 - y^2 & \text{if } x^2 + y^2 < R^2 \\ 0 & \text{if } x^2 + y^2 \geq R^2 \end{cases}$$

This model is based on the assumption that a mass in a mammogram can be approximated as the projection of a sphere, which will be made plausible in Section 4.3. It appeared that the ratio between the radius of the the mass (R) and the radius of the window should be chosen with care. The value of 1.3 was chosen, because for this ratio the variance of the template for this model function reaches its maximum value, and the best results are obtained.

For each pixel in the mammogram, the correlation between the template T and the area around the pixel covered by the template window S was computed with

$$Cor(T,S) = \frac{Cov(T,S)}{\sqrt{Var(T)Var(S)}},$$

where $Cov(T,S)$ is the covariance between the template T and the region S of the mammogram, and $Var(S)$ and $Var(T)$ are respectively the variance of the mammogram inside the template window and the variance of the template. The covariance between the template and the region S can be computed by

$$Cov(T,S) = \frac{1}{N_S} \sum_{(x,y)\in S} (I(x,y) - E(S))(T(x,y) - E(T))$$

where N_S is the number of pixels in the window, and $E(S)$ and $E(T)$ are respectively the average values of the image inside the window and the template.

The correlation value for each pixel is between -1 and 1, where a high value indicates the possible presence of a mass. In contrast to the convolution based method described in the previous section, this correlation value is independent of a linear scaling of the intensity values. Sometimes template matching is implemented by computing the covariance instead of the correlation. However, by using the correlation instead of the covariance, faint masses may be detected that only have small covariance values. An example feature image can be seen in Figure 4.1c.

The scale parameter for template matching is the radius of the mass in the template.

4.2.3 Gradient orientation analysis

The gradient orientation analysis method is an adaptation of a method for detecting stellate lesions that was described in [10]. Masses appear in mammograms as more or less circular bright regions. Therefore, a mass would appear in a map of gradient orientations as a circular region with many gradients pointing towards the center. These regions can be detected by statistical analysis of the gradient orientation map.

In Figure 4.2 the method is explained. A measure of suspiciousness for the pixel in the center i is computed. For all pixels j with at a distance between *rmin* and *rmax* of i, it is determined whether the gradient orientation vector is oriented towards the center. A pixel is considered to be oriented towards the center if the angle θ between the line through i and

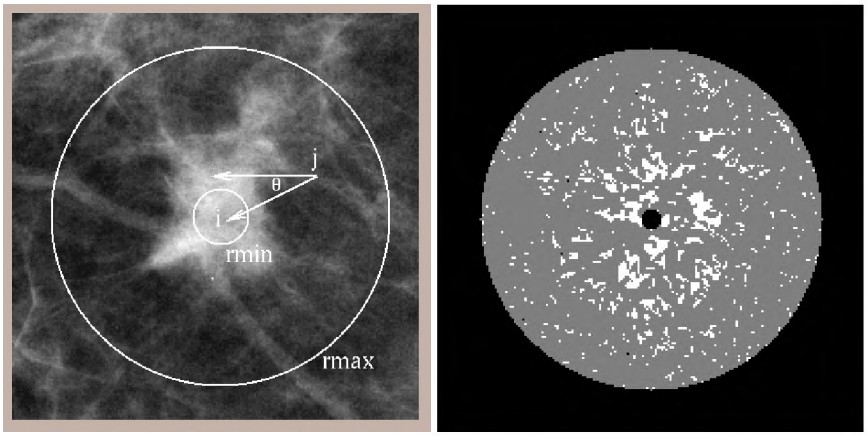

Figure 4.2: Explanation of the gradient orientation method. (a) Circular neighborhood around pixel i. For pixel j, the orientation of the gradient is given by the top arrow, the bottom arrow shows the orientation towards the center. (b) The grey and white points are in the circular neighborhood of pixel i. White points are oriented towards the center, grey points are not.

j and the gradient orientation is smaller than a constant D divided by the distance between i and j. A small D will reduce the number of neighbor pixels that is oriented towards the center. In Figure 4.2b all grey and white pixels are in the circular neighborhood, where the white pixels are oriented towards the center.

If no structure is present, the probability p_j that a pixel j is randomly pointing towards the center i is given by

$$p_j = \frac{D}{d_{ij}\pi},$$

where d_{ij} is the distance between i and j. The mean probability p that a pixel in the neighborhood is oriented towards the center is

$$p = \frac{1}{N}\sum_j \frac{D}{d_{ij}\pi},$$

where N is the number of pixels in the circular neighborhood. Therefore, for each circular neighborhood we can compute the expected number of points with an orientation towards the center Np, and its variance $\sqrt{Np(1-p)}$. The number of pixels n that is actually pointing towards the center is counted, and the feature representing suspiciousness is computed by

$$\text{Gradient feature} = \frac{n - Np}{\sqrt{Np(1-p)}}.$$

This feature is adaptive to varying neighborhood sizes and D, and does not suffer from artifacts near the edge of the breast, where the circular neighborhood lies partly outside the breast area.

The gradient orientation method depends on accurate calculation of the orientation map, and on the spatial scale at which the derivatives are computed. In our implementation, first order Gaussian derivative were used to compute the gradients in two orthogonal directions

I_x and I_y, where the scale of the Gaussian determines the resolution. For each pixel in the image, the gradient orientation is computed with

$$\theta(x,y) = \tan^{-1}(\frac{I_y}{I_x}),$$

yielding the gradient orientation image map.

A number of parameters have to be chosen: the size of the neighborhood that is determined by *rmin* and *rmax*, the resolution at which the gradients are computed, and the parameter D that determines whether or not a gradient is oriented towards the central pixel. An example of a mammogram with its feature image for this feature can be seen in Figure 4.1a and 4.1d.

The scale parameter is the gradient orientation approach is *rmax*. In Figure 4.2, the *rmax* is chosen rather large. If the mass is embedded in a fatty area, this does not strongly affect the response of the method, but in a more dense area the signal yielded by the mass might get lost by taking *rmax* too large.

4.2.4 Single and multi-scale detection of masses

Because of the wide range of sizes masses can have, a multi-scale approach may be fruitful. All methods described above can be made multi-scale by applying the method for each pixel on a number of scales, and combining these in some way. For the gradient orientation method, this means varying *rmax*, for the Laplacian varying sigma, and for the template matching the radius of the model of the mass R.

In this work, two approaches to use the methods in a multi-scale manner were applied. The first is providing the output of a detection method over a range of scales to a classification scheme. In this work a neural network was used to map these features to a single combined measure of suspiciousness, a value between 0 and 1. The neural network used was a simple 3 layer feed-forward network with 5 hidden nodes, trained with the back-propagation algorithm. A set of 63 mammograms containing malignant tumors were used to select normal and abnormal training patterns for training the neural network. These mammograms were not used for test purposes. The second approach to do multi-scale detection was taking the maximum value over all the scales. The template matching method and the gradient orientation method have a normalized output and scales can be combined without problems, but this is not the case for the Laplacian filtering method. Using the maximum value over the scales may therefore be suboptimal.

The multi-scale approach for the gradient orientation method can be computed without much overhead compared to the a single-scale approach. The template matching scheme requires more computation, because at each scale the filter output has to be computed separately, and combined afterwards. The Laplacian filter also requires extra computation, but when implemented as a DoG-filter, each Gaussian convolution can be used twice. (of course, this restricts the choice of the scales to scales differing with a factor close to 1.6).

4.3 Experiments on simulated masses

To examine the sensitivity of the performance of the three methods for variations of the scale parameter, simulated masses were used. Simulated masses can be easily altered, and

the effect on the performance of the three methods on variations in size and brightness of the mass can be examined in a quantitative way.

It is not our goal to compare the various methods based on the results of the simulations. Such a comparison would not be valid because it may be the case that simplifications in the way tumors are simulated may be favorable for one of the methods and not for the others. Our main intention is to examine the changes of the response of our methods on variations of the size of masses, and to find appropriate values for the scale parameters. The actual comparison of the detection methods will be performed on a large database of mammograms.

4.3.1 Simulation method

A mass was modeled as a sphere. Because tumor tissue is relatively hard, it was assumed that the effects of compression on the shape can be neglected in a first approximation. The following equation describes the relation between the exposure and the absorption of the various types of tissue present

$$E = E_0 \exp(-\sum_i (\mu_i d_i)).$$

The μ_i's are the linear attenuation coefficients of tissues like fat, glandular tissue and tumor tissue, the corresponding d_i's represent the thickness of these tissues. This model does not take into account the effects of scatter, the divergence of the X-ray beam and the anode heel effect, and therefore is a simplification of the real process. In the linear part of the film curve, the optical density OD of the exposed film is given by

$$OD = c_1 \log(E) + c_2.$$

Using a calibrated image digitizer with a linear relation between pixel intensity and optical density, the pixel intensity value y becomes

$$y = a + b \sum_i (\mu_i d_i),$$

with a and b constants. A simulated mass is projected on regions with normal mammographic tissue. It is assumed that the mass replaces a homogeneous spherical volume of breast tissue. Because the pixel intensity is linear in the thickness of the tumor we want to simulate, we can just add the term

$$L(x,y) = B * \sqrt{\max(R^2 - (x-m)^2 - (y-n)^2, 0)}$$

to a mammographic image to create a mass with a radius R at location (m,n). By altering the brightness B and the radius R the brightness and size of the mass can be varied. An example of a mass projected on a region of normal tissue is shown in Figure 4.3.

To generate a realistic sample of masses for the simulations, the size distribution of masses in mammograms should be used. Based on 314 annotation of masses in several datasets, made by an experienced radiologist, the size distribution of masses was approximated. Only the mass was annotated, if spicules were present they were left outside the

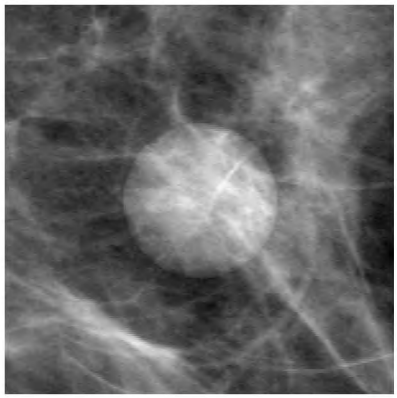

Figure 4.3: A simulated mass embedded in mammographic tissue.

annotated region. The distribution of the size of the annotations is shown in Figure 4.4a. According to the size distribution, masses with a radius between 2 and 18 millimeter were simulated.

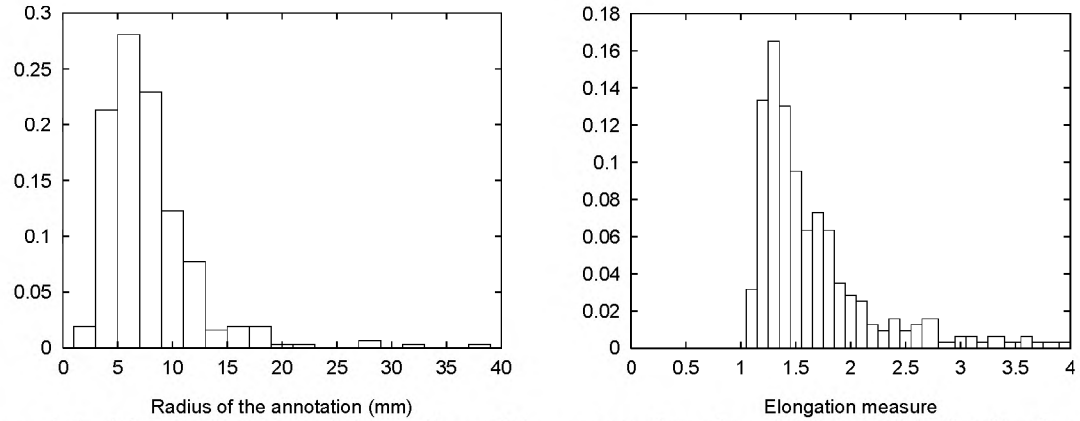

Figure 4.4: (a) Distribution of the radius of the annotations in millimeter.(b)Distribution of the elongation measure.

Using these annotations, our assumption that the projections of masses are circular objects was justified. For each annotation the center of gravity was determined. Its elongation was computed using the minimal and maximal distance of the edge from the center of gravity, R_{min} and R_{max}, with

$$\text{Elongation} = \frac{R_{max}}{R_{min}}$$

The distribution of this elongation measure is shown in Figure 4.4b. Most masses have a measure below 2, and many of them even below 1.5. Although only a few masses are really circular, most of them have an elongation measure that is reasonable close. This result raises our confidence that for our scale experiments it is not too unrealistic to simulate masses as circular objects. No relation was found between between the size of the annotation and the elongation.

4.3.2 Simulation experiments

From 26 mammograms without abnormalities, 168 regions with normal mammographic tissue of 6.25x6.25cm were extracted. Both fatty and dense regions were present in this set. On each region, 9 masses with a radius ranging from 2 mm to 18 mm with 2 mm steps were projected in the way described in the previous section. For each size mass, this yielded a set of 168 different regions.

To examine the performance of the three mass detection methods at various scales, they were applied on 8 different scale levels to all 9 sets of 168 mass projections. For each scale, this gave a feature value distribution for pixels inside masses of a specific size. For each mass, only the feature value at the pixel in the center was used to compute the distribution.

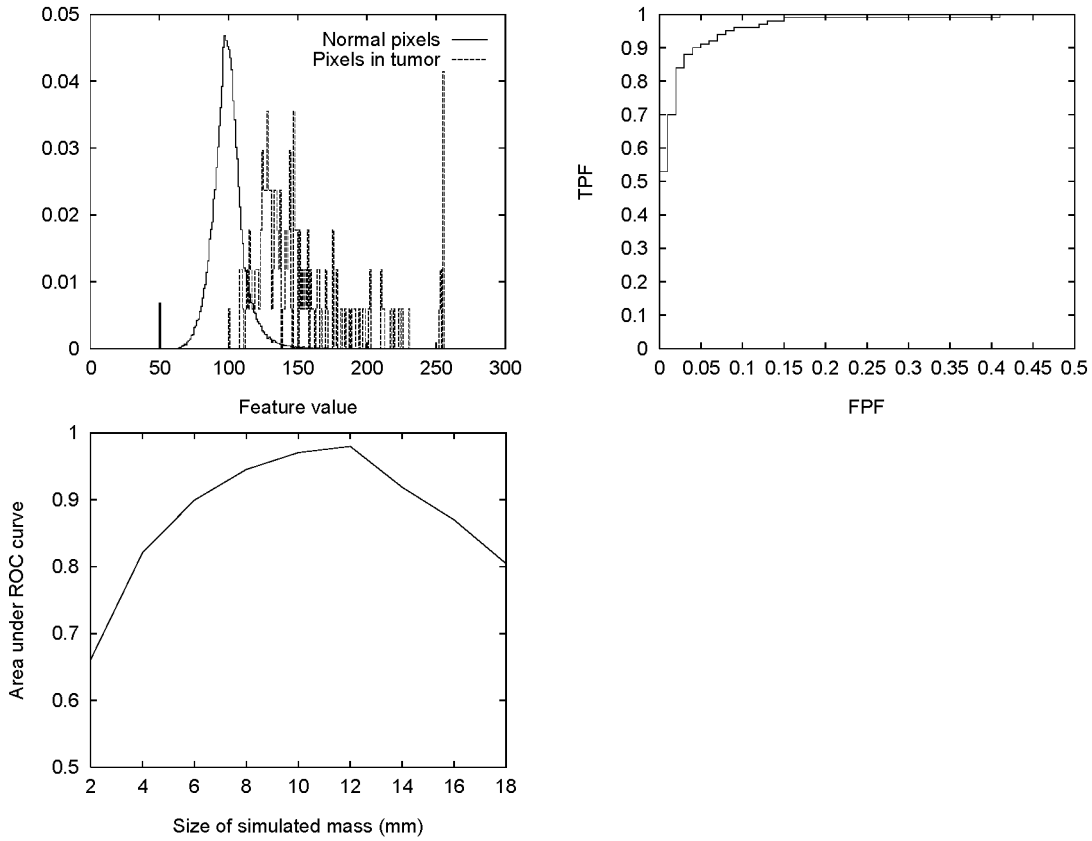

Figure 4.5: Performance measurement for the simulations. The gradient orientation method was used with *rmax*=12 mm, the radius of the simulated masses was also 12 mm. (a) Distribution of feature values of pixels in normal tissue and of the center pixel of the simulated masses. (b) The corresponding ROC curve. (c) Area under the ROC curve as a function of the size of the simulated mass.

Each method was also applied on the same scales to 10 normal mammograms, generating a distribution for pixels in normal tissue. As an example, in Figure 4.5a, the two distributions for the gradient orientation analysis method with a scale parameter of 12 mm are shown. For the distribution for the masses, masses were used with a size of 12 mm. For the multi-scale analysis, the output of the neural network was used to compute the distributions for normal pixels and the central pixel of the simulated masses.

Pixels in the center of a simulated mass have higher values than pixels in normal tissue,

but the two distributions partly overlap. Due to the large amount of normal pixels present in the 10 mammograms, the normal distribution is very smooth. The distribution of the values of the central pixel of the simulated tumors is much less smooth, due to the limited number of 168 simulated masses. Based on the two distributions ROC curves were determined, as shown in Figure 4.5b.

As a measure for detectability of a tumor of a given size, the A_z-value (the area under the ROC curve) was computed. This way, the performance of the methods versus the size of the simulated masses could be examined, which is shown in Figure 4.5c. An A_z value of 1 is a perfect score, a value of 0.5 is achieved when no information is present in a feature. The example shows that the gradient orientation method with its scale parameter set to 12 mm has optimal performance for masses of that size, and that the performance degrades for larger and smaller masses.

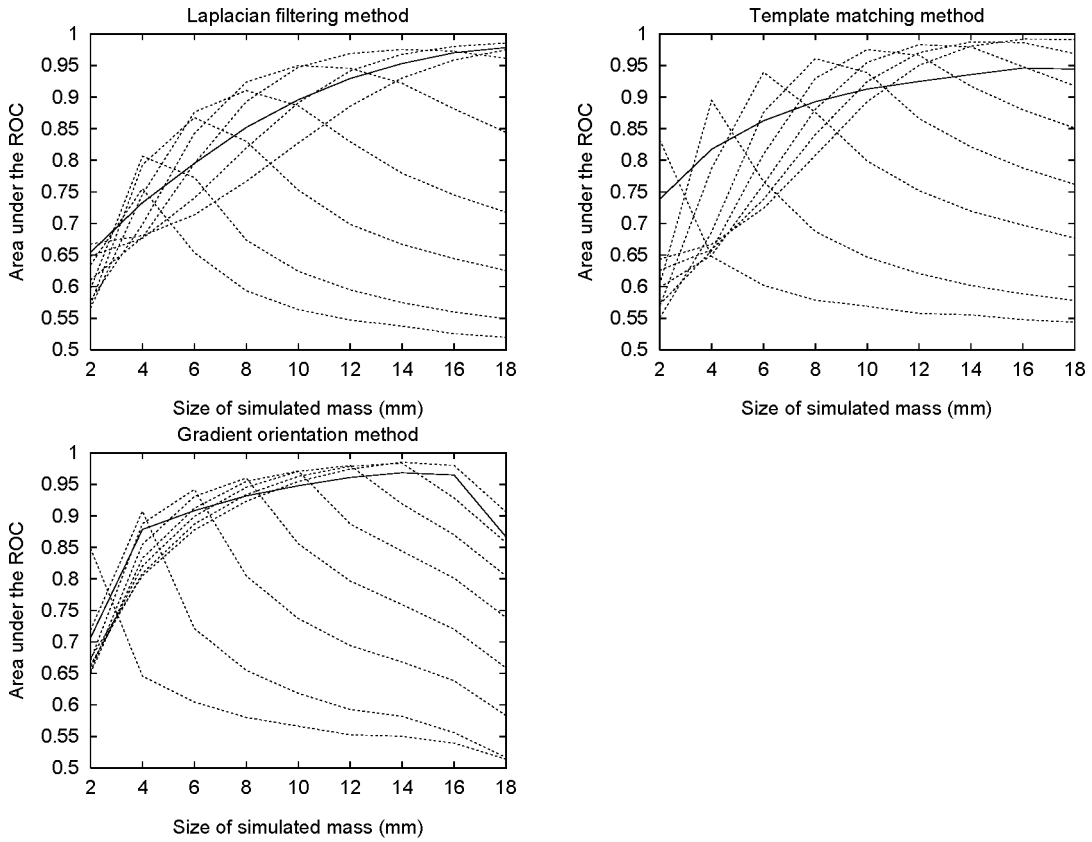

Figure 4.6: Results on simulated masses. The dashed lines show the performance of the method at various single scales, the solid lines the results for the max-scale method. (a) Laplacian filtering. (b) Template matching. (c) The gradient orientation feature.

In Figure 4.6 the performances of all three methods are shown. Each figure shows the results of one of the methods for a number of scales. Horizontally the radius of the simulated mass is shown. On the vertical axis, the area under the ROC curve is given. Each dashed line gives the results of a method applied at a different scale. The results show that the template matching scheme has a very good performance for masses of the size that correspond exactly to the scale of the model, but has a rather strong decrease for masses of a different size. The two other methods have a wider range of sizes with good performance. The Laplacian filtering approach has considerable lower peaks than the other two methods.

The solid line represent the performance of the method applied multi-scale using the maximum value approach. The maximum scale performed better than the neural network. The output of the neural network was similar to the output on a medium scale with a somewhat wider performance range, but less than obtained with the maximum value approach. This is due to the large number of medium size tumors, and a relatively small number of small and large tumors in the set. In Figure 4.7 the results of both multi-scale approaches and a single-scale result are shown for the template matching method. Experiments were done to determine the optimal number of scales that were used with the neural network, but little difference was found between the output when 8 scales were used compared to when only 3 scales were used. For a specific size, the multi-scale methods are outperformed by the single-scale method applied at the optimal scale, but the multi-scale methods perform well over a wider range of sizes. If the size of masses would be uniformly distributed over this range, multi-scale detection would be superior.

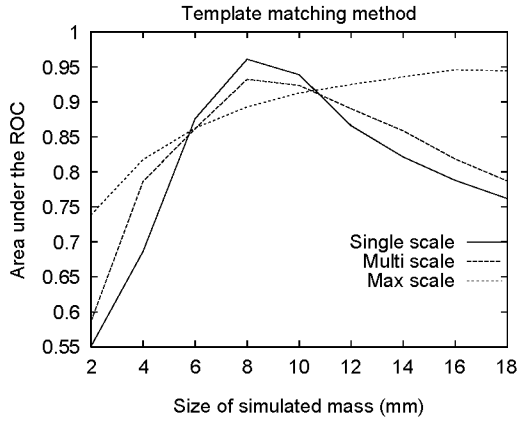

Figure 4.7: Single and multi-scale approaches compared for the template matching method. The multi-scale methods have a wider good performance range, but a lower top than the single-scale approach.

When masses were simulated with varying size according to the distribution as determined from our database, scale parameters that yield good overall performance on a sample of masses with sizes that are found in a breast cancer screening were determined. The overall performance for each method was computed directly from the A_z values for the various sizes by weighting them according to the distribution. For each scale, the area under the ROC curve was computed, and the results are shown in Figure 4.8. For example, template matching was optimal when the model had a radius of 8 mm, for which an area under the ROC was achieved of 0.85. The optimal values are also shown in Table 4.1. For the gradient orientation method, not much gain was achieved by applying the detection method on several scales. For the template matching method, the A_z value increased slightly from 0.85 to 0.87. Although the Laplacian convolution method performed worse for masses that correspond to the filter scale, in a single-scale approach it had a similar overall performance as the template matching model due to its wide good performance range. Multi-scale analysis using the neural network yielded similar results, but when the maximum value was used, performance was lower due to the lack of normalization. Multi-scale analysis was most useful for the template method which has high peaks but a limited good performance range.

The gradient orientation method contains a few other parameters aside from the scale parameter *rmax*. The chosen values for these parameters appeared to be much less critical.

Method	Single scale		Multi scale	Maximum scale
	Scale	A_z	A_z	A_z
Laplacian convolution	sigma=5.4	0.85	0.85	0.82
Template matching	R=8	0.85	0.87	0.87
Gradient orientation	rmax=12	0.90	0.90	0.91

Table 4.1: Optimal scales and corresponding A_z values for the three detection methods when the masses are simulated according to the distribution of Figure 4.4. Both for the single-scale approach and the two multi-scale approaches the results are given.

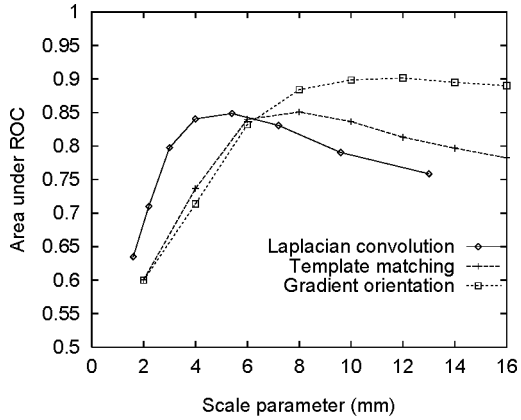

Figure 4.8: The relation between the scale parameter and the performance on masses simulated according to the size distribution shown in Figure 4.4. Horizontally the scale parameter is shown, which is the radius for the gradient orientation methods and the template matching filter and the sigma for the Laplacian, vertically the corresponding area under the ROC curve is given.

Both D and the resolution for the computation of the gradient orientation were varied. A high resolution gave best results, but results only slightly degraded for lower resolutions. A sigma of 0.2 mm was used in our experiments, smaller values require a smaller pixel size to be computed accurately. However, we do not expect any gain in using smaller sigmas, as the differences between results obtained with sigmas between 0.5 mm to 5 mm were very small. The value for D proved to be even less critical. Very little change was found in the results for D varying between 2 mm and 10 mm.

4.4 Experiments on mammograms

To compare the performance of the three mass detection methods, a database of mammograms of 71 cases with a malignant tumor was used. All mammograms were digitized at a resolution of 50 μm per pixel with a model 85 Lumisys digitizer, after which the resolution was averaged down to 200 μm per pixel.

To decrease the computational load, images were processed in a sampling mode, where features were computed at regularly spaced test locations. A grid with an interval of 8 pixels (1.6mm) was used, which appears to be sufficiently dense to avoid missing small masses. The features were computed at these locations using the full resolution of 200 μm per pixel.

The feature images obtained were thresholded at various levels. At each threshold level a number of regions are segmented, which are considered to be suspicious. By varying the threshold level, sensitivity and specificity of the detection method can be altered. Low threshold levels will give high sensitivity, but many false positive regions will be signaled. High threshold values will reduce the number of false positives, but lower the sensitivity. FROC studies were carried out to show the relation between the sensitivity and number of false positives. On the horizontal axis, the number of false positives is shown, vertically the corresponding sensitivity (percentage of detected tumors). For the multi-scale experiments with the neural network, the likelihood-image that is yielded by the output of the neural network was used instead of a feature image, but the rest of the procedure is exactly the same.

A tumor was considered detected if the pixel with the highest measure of suspiciousness in the segmented area was located inside the annotation. If this peak was not lying in an annotated region, this area was considered to be a false positive. If a peak is found closer than 1cm from another peak, it was considered to belong to the same suspicious region and the lowest was removed.

4.4.1 Data set

The dataset used is a consecutive set of 71 masses that appeared between 1993 and 1996 in screening in Nijmegen. Only cases in which the only sign of malignancy was a cluster of microcalcifications were not included in this set. All women participating in this program are between 50 and 69 years old. In this screening program, if a woman is screened for the first time both oblique and cranio-caudal films are made. On succeeding visits, cranio-caudal films are made when the radiographers find the oblique films hard to read due to dense tissue, or when they find a suspicious area. For each case in the database, the oblique films were present, in 61 cases cranio-caudal films were also available. This makes a total of 132 mammograms with a visible malignant tumor, ranging from very subtle to very obvious, but a typical sample of tumors that occur in screening.

4.4.2 Experiments

All three detection methods were applied to the test set of 132 mammograms. The results are shown in Figure 4.9. As was predicted by our simulations, the Laplacian convolution method performs worse than the other two, due to the strong dependence on the brightness instead of shape. Because of these results, we did not investigate this method in more detail but focused on the gradient orientation method and template matching method.

The test set of 132 mammograms was divided in three sets: 31 mammograms with a small tumor (radius smaller than 6mm), 63 with a medium size tumor (radius between 6 mm and 10 mm) and 38 with a large tumor (radius larger than 10mm). Note that in this work the radius is used for the size of masses, nor the diameter. The mass detection methods were applied on a small, medium and large scale (for the template matching method respectively an R of 4, 8 and 12 mm, for the gradient orientation an *Rmax* of 4, 8 and 12 mm) to each of these subsets, and to the total set of 132 mammograms. For the template matching method, the results are shown in Figure 4.10a-d. In the first figure, the results are shown for small tumors. A small scale value gives optimal results. In Figure 4.10b and 4.10c, the results are

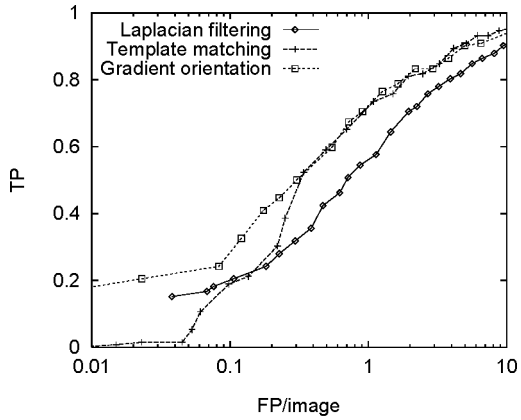

Figure 4.9: Results on the dataset for the three detection methods applied at a medium scale.

shown on medium and large tumors, where a corresponding scale value gives best results. In Figure 4.10d the results are shown on the total set of 132 mammograms. The same analysis was done for the gradient orientation method and similar behavior was found, although the influence of the scale parameter was less strong (Figure 4.11). For this method, a medium scale value had reasonable performance over a fairly large range of sizes.

The use of multi-scale detection was also investigated and the results were compared to the results of the best single-scale in Figure 4.12. For the gradient orientation analysis no improvement was achieved, and the curves obtained were very similar to the curve obtained using the optimal single scale. The template matching method did benefit from the multi-scale approach when the maximum value was used, especially at lower specificity levels where a slightly higher sensitivity was achieved compared to the single-scale approach. At higher specificity levels, no difference was found.

4.5 Discussion

The experiments on simulated masses showed that scale is an important parameter. Figure 4.8 shows the relation between the chosen scale and the performance of the three methods. The template matching method and the Laplacian convolution method have a small interval for the scale parameter where the method is optimal. For the gradient orientation method, choosing the optimal scale is less critical. On the simulated masses, the gradient orientation method performed better than the other two methods (Table 4.1).

Figures 4.10 and 4.11 show that by varying the scale parameter the performance of the methods for small, medium or large masses changes. If the scale parameter is set to small, performance on large masses becomes worse, if it is too large small masses will be missed more often. A good value for the scale parameter will make the method capable of detecting masses in a range of sizes that occurs in screening. The template matching method has very good performance for masses with a size close to the corresponding scale, especially for small and medium size tumors, but the performance on other sizes is much worse. Figure 4.11 shows that at a medium scale, the performance of the gradient orientation analysis method is almost as good as that of a small scale for small tumors and at a large scale for large tumors. Therefore, not much gain is achieved by a multi-scale approach, which is in agreement with Table 4.1 which shows that the area under the ROC curve for the multi-

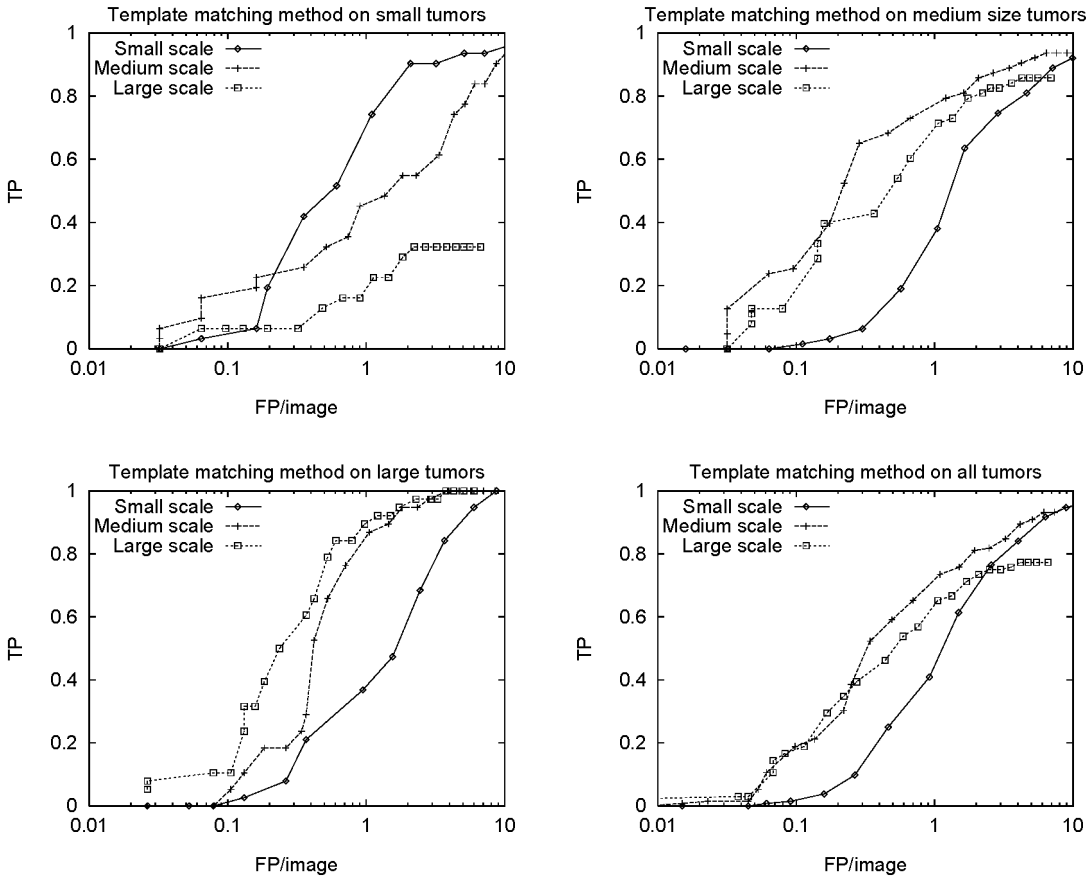

Figure 4.10: Results for the template matching approach on three scales. (a) Small tumors. (b) Medium size tumors. (c) Large tumors. (d) All tumors.

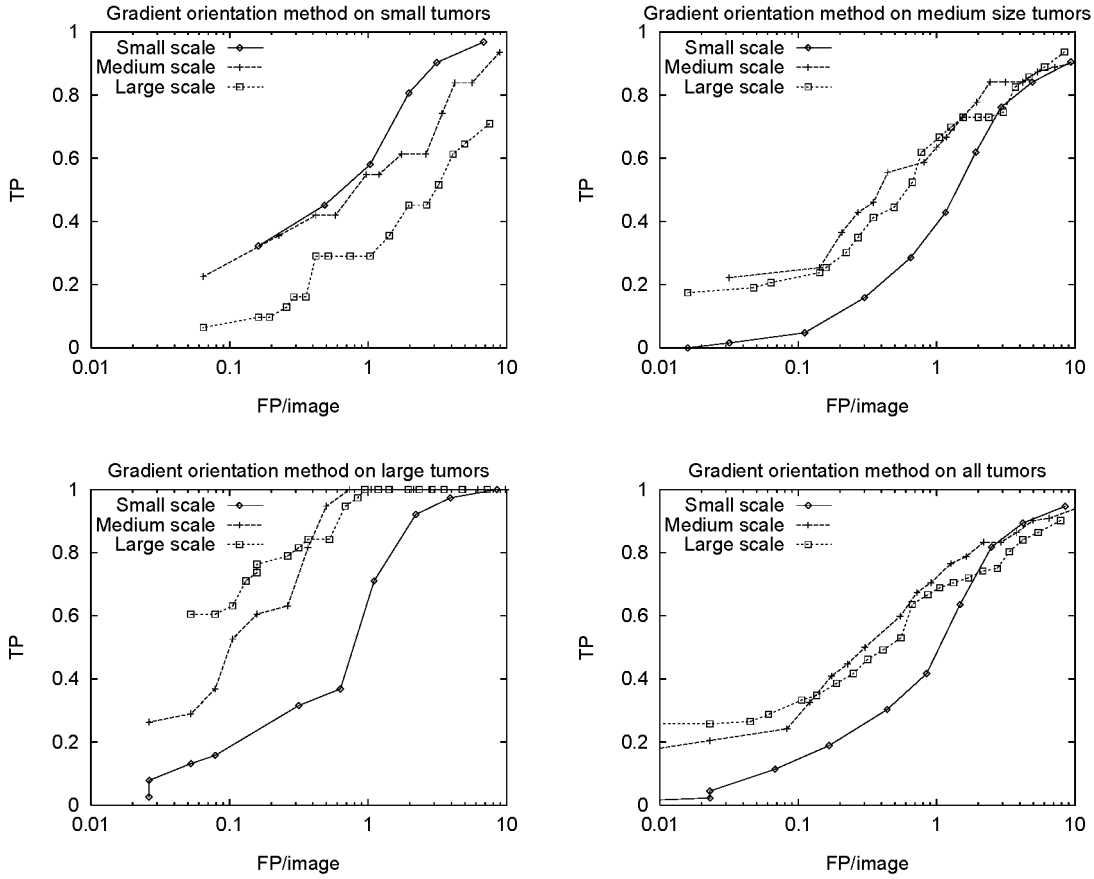

Figure 4.11: Results for the gradient orientation approach on three scales. (a) Small tumors. (b) Medium size tumors. (c) Large tumors. (d) All tumors.

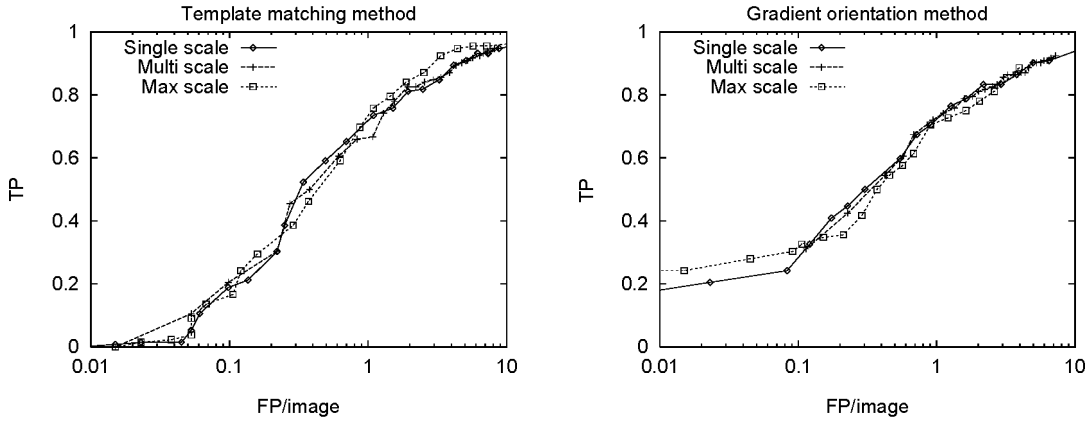

Figure 4.12: Results on the mammograms for the single and the two multi-scale approaches of the template matching method and the gradient orientation method.

scale approaches is almost equal to the area under the ROC curve for the optimal single-scale approach. For the template matching method, a matching scale is more beneficial, and therefore a multi-scale approach is more useful, which is shown in Figure 4.12. The template matching scheme benefits from a multi-scale approach, especially at lower specificity levels when the maximum value is used. The multi-scale curve for the gradient orientation method is very similar to its single-scale curve.

The optimal single-scale approach is good for average sized tumors, missing mainly small masses, where the multi-scale approach performs equally well over all sizes. A disadvantage of the multi-scale approach is that both small and large normal tissue structures generate additional false positives in comparison to a single-scale approach. There, only normal structures of the corresponding scale generate false positives. With multi-scale detection, both small and large tumors give strong signals, but due to the extra false positives the FROC curve is comparable to the single-scale curves. However, it may be the case that a second step to remove false positives is more successful if all tumors give a strong signal. Regional analysis using texture measures etcetera for all suspicious regions may result in the removal of many false positives.

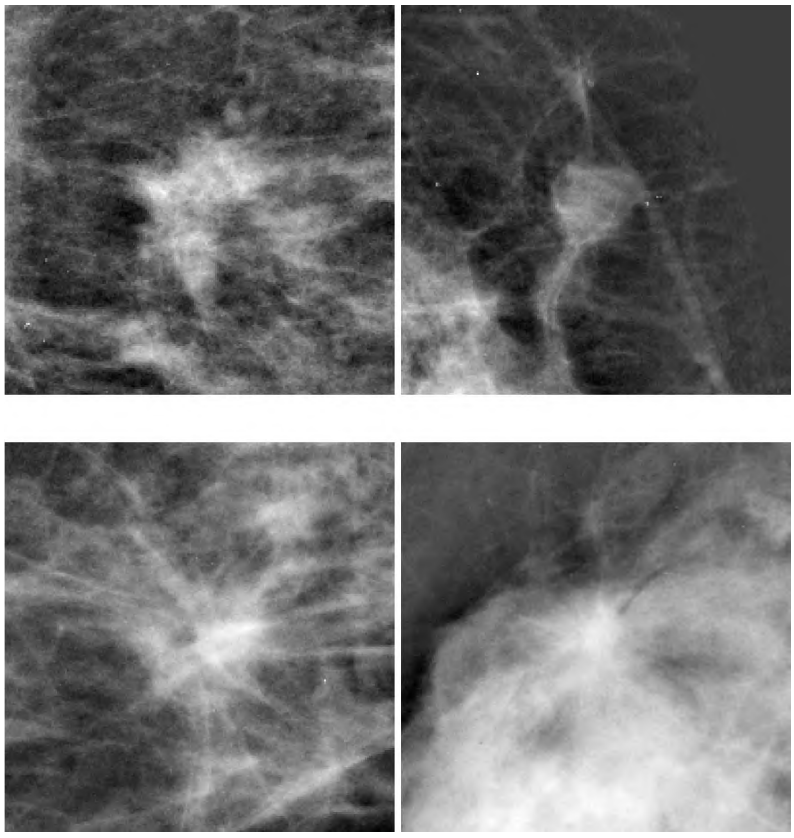

Figure 4.13: Four examples of masses. (a),(b),(c) Cases for which the template matching method outperforms the gradient orientation method. (d) Case for which the gradient orientation method outperforms the template matching method.

The main difference between the results on the mammograms and on the simulated masses is the performance of the template matching method, which has equal performance on real mammograms as the gradient orientation method. There are a number of reasons

for this phenomenon. The template matching method is less distracted by spicules, and performs better on masses that are not really circular. Especially masses in a fatty area are found at better specificity levels. Four typical cases are shown in Figure 4.13. In Figure 4.13a, the mass is not really circular, making the gradient orientation method not very successful. The mass was only found at a specificity level of over 4 false positives per image by this method. However, due to the fatty surroundings, the template method detected the tumor at 0.6 false positives per image. The mass in Figure 4.13b was found by both methods, but at a much better specificity level by the template method, again due to the fatty surroundings (0.8 versus 2 false positives per image). The mass in Figure 4.13c shows clear spiculation, which reduced the response for the gradient orientation method. This was not the case for the template matching method (0.2 versus 2 false positives per image). Figure 4.13d shows an example where the gradient orientation method outperformed the template matching method. The surrounding area has no influence on the gradient orientation method, but confused the template matching scheme (0.7 versus 1.9 false positives per image). The Laplacian convolution method performed worse than the other two methods on real masses, as is shown in Figure 4.9. Especially subtle masses with low contrast are difficult to detect for this method.

It seems that none of these methods has a sensitivity high enough for an initial detection method where a close to 100% sensitivity is desired. It should be kept in mind however, that the database contained consecutive cases. For some cases, the primary sign was spiculation or asymmetry or an architectural distortion. Features to detect these signs can be added in the detection scheme to obtain high sensitivity [11].

4.6 Conclusions

Simulated masses can be used to examine the behavior of mass detection methods for changes in size and intensity of the masses. Changes in performance of the methods for varying parameter values can be examined in a quantitative way. The results of the methods on simulated masses agree to a large extent to real masses, which gives confidence in the way the masses are simulated.

Choosing a correct scale for optimal detection is important. The detection methods have worse performance when the scale is chosen suboptimal. No strong improvement was found when the detection was performed in a multi-scale way for any of the three methods that were examined in this work. Multi-scale detection is most useful for methods which selectively respond to masses in a small range of sizes, which was only the case for template matching. This method shows a slight improvement when detection is performed on multiple scales, especially at low specificity levels. It should be kept in mind that all methods were pixel-based methods, region-based methods may benefit more from multi-scale detection.

The Laplacian convolution method is not very suited for mass detection, due to its strong dependence on intensity instead of shape. The template matching scheme and the gradient orientation analysis method performed better, both in single- and multi-scale schemes. Similar FROC curves were obtained for these two methods when applied to a database of real mammograms.

Bibliography

[1] S Astley, I Hutt, S Adamson, P Rose, P Miller, C Boggis, C Taylor, T Valentine, and J Davies. Automation in mammography: computer vision and human perception. *SPIE 1905*, pages 716–730, 1993.

[2] C J Baines, D V McFarlane, and A B Miller. The role of the reference radiologist: estimates of inter-observer agreement and potential delay in cancer detection in the national breast screening study. *Inv Radiol*, 25:971–976, 1990.

[3] R E Bird, T W Wallace, and B C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184:613–617, 1992.

[4] D Brzakovic and M Neskovic. Mammogram screening using multiresolution-based image segmentation. In K W Bowyer and S M Astley, editors, *State of the art in digital mammographic image analysis*, volume 9 of *Series in machine perception and artificial intelligence*, pages 103–127. World Scientific, 1994.

[5] H P Chan, K Doi, C J Vyborny, R A Schmidt, C E Metz, K L Lam, T Ogura, Y Wu, and H Macmahon. Improvement in radiologist's detection of clustered microcalcifications on mammograms. *Inv Radiol*, 25:1102–1110, 1990.

[6] B R Groshong and W P Kegelmeyer. Evaluation of a hough transform method for circumscribed lesion detection. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 361–366. Elsevier, Amsterdam, 1996.

[7] J E Harvey, L L Fajardo, and C A Inis. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. *AJR*, 161:1167–1172, 1993.

[8] I W Hutt, S M Astley, and C R M Boggis. Prompting as an aid to diagnosis in mammography. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 389–398. Elsevier, Amsterdam, 1994.

[9] N Karssemeijer. Recognition of stellate lesions in digital mammograms. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 211–220. Elsevier, Amsterdam, 1994.

[10] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[11] N Karssemeijer and GM te Brake. Combining single view features and asymmetry for detection of mass lesions. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 95–102. Kluwer, Dordrecht, 1998.

[12] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[13] S L Kok, J M Brady, and L Tarrasenko. The detection of abnormalities in mammograms. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 261–270. Elsevier, Amsterdam, 1994.

[14] S M Lai, X Li, and W F Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans on Med Imag*, 8:377–386, 1989.

[15] L Li, W Qian, and L P Clarke. Digital mammography: computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms. *Academic radiology*, 11(4):724–731, 1997.

[16] L Miller and N Ramsey. The detection of malignant masses by non-linear multiscale analysis. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 335–340. Elsevier, Amsterdam, 1996.

[17] S L Ng and W F Bischof. Automated detection and classification of breast tumors. *Comput Biomed Res*, 25:218–237, 1992.

[18] R M Nishikawa, M L Giger, K Doi, C J Vyborny, and R A Schmidt. Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms. In K W Bowyer and S M Astley, editors, *State of the art in digital mammographic image analysis*, volume 9 of *Series in machine perception and artificial intelligence*, pages 82–102. World Scientific, 1994.

[19] B Sahiner, H P Chan, N Petrick, D Wei, M A Helvie, D D Adler, and M M Goodsitt. Classification of mass and normal breast tissue : a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imag*, 15:598–610, 10 1996.

[20] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[21] G M te Brake and N Karssemeijer. Automated detection of breast carcinomas not detected in a screening program. *Radiology*, 207:465–471, 1998.

[22] J A M van Dijck, L M Verbeek, Hendriks J H C L, and R Holland. The current detectability of breast cancer in a mammographic screening program. *Cancer*, 72:1933–1938, 1993.

[23] T Y Young and K S FU (eds). *Handbook of pattern recognition and image processing*. Academic Press Inc., London, UK, 1986.

[24] B Zheng, Y H Chang, and D Gur. Computerized detection of masses in digitized mammograms using single image segmentation and a multi-layer topographic feature analysis. *Acad Radiol*, 2:959–966, 1995.

[25] R Zwiggelaar, J E Schumm, and C J Taylor. The detection of abnormal masses in mammograms. In *Medical Image Understanding and Analysis*, University of Oxford, UK, July 7-8 1997.

Chapter 5

A discrete dynamic contour model for mass segmentation[1]

5.1 Introduction

Many research groups have developed pixel-based algorithms to detect masses in digital mammograms [15, 2, 3]. These algorithms are generally very sensitive but signal many false positive regions per image. Our approach to detect masses aims at detecting both spicules and the central mass [2, 13, 14]. The presence of either one or both of these properties can trigger the system to signal a suspicious region. When a candidate mass is found, the initial pixel-based step can be followed by a region-based step to examine the suspicious areas more closely. Goal of this second step is to reject false positive regions to increase the specificity while maintaining high sensitivity. An important part of this step is the segmentation of the suspect region into background tissue and the region suspect for being a mass. Based on this segmentation, shape and contrast features can be computed and an examination of the edge of the mass is possible, providing information that can be used to classify the region into one of the classes normal, benign or malignant.

A large number of segmentation methods have been developed in the field of image analysis and many of them have been used to segment masses in mammograms. Petrick et al. [9] used a density-weighted contrast enhancement segmentation method. Markov random fields were used by Comer et al. [1] and Li et al. [5] to segment regions based on texture information. A segmentation method based on fuzzy partitioning was developed by Sameti et al. [11]. One of the most popular segmentation methods, used in many image processing fields, is region growing. Region growing has been applied to segment masses by a number of groups [5, 10, 8, 4]. In recent years, deformable models have become popular in the field of medical image analysis [7]. We have applied a member of this family, a discrete dynamic contour model, to the task of mass segmentation. The implementation of the discrete dynamic contour model was based on an algorithm described by Lobregt and Viergever [6], which is a fast and robust procedure to detect boundaries. To get insight in the performance of this model, it was compared to the region growing method described by Kupinski and Giger [4].

[1]Published as: G.M. te Brake, M.J. Stoutjesdijk, N. Karssemeijer, *A discrete Dynamic Contour Model for Mass Segmentation in Digital Mammograms*, SPIE Medical Imaging 1999, Vol. 3661, pp 911-919, 1999.

To examine the strengths and weaknesses of the methods, two experiments were done. Both segmentation methods need a starting point. In the first experiment, for each mass the center of gravity of the annotation was used. Generally, these points yield good results, and therefore the maximum performance of the methods can be determined, as well as their sensitivity to a number of internal parameters. In the second experiment, the pixel-based initial detection step was used to generate starting points. These starting points are often not as good, and robustness of the method to this sub-optimality was examined.

The success of a false positive removal step strongly depends on the synergy between the segmentation step and the classification step. The quality of a segmentation should therefore be examined based on the effectiveness of the features that use this segmentation. However, as a first step it is common practice to compare the segmented regions with the annotations that have been made by the radiologist, because strong correlation in the performance between these two performance measurements can be expected. This is also the way evaluation was done in this work.

5.2 A Discrete dynamic contour model

Our implementation of the discrete contour model is based on an algorithm developed by Lobregt and Viergever [6], which is a fast and robust procedure to detect the boundary of a region. Unlike snakes, it is a discrete model represented by vertices that are connected by edges. An initial contour has to be chosen, after which each vertex is moved around by a combination of internal and external forces working on it. These forces determine the speed and acceleration of the vertex.

The model consists of a number of vertices connected by edges. For each vertex i with connecting edges d_i and d_{i-1}, a local coordinate system is constructed represented by a tangential unit vector $\hat{t}_i$ and a radial unit vector $\hat{r}_i$

$$\hat{t}_i = \frac{\hat{d}_i + \hat{d}_{i-1}}{||\hat{d}_i + \hat{d}_{i-1}||}$$

and

$$\hat{r}_i = \begin{vmatrix} 0 & 1 \\ -1 & 0 \end{vmatrix} \hat{t}_i,$$

where $\hat{d}_i$ denotes the unit vector of d_i. The tangential vector is "in line" with the contour, the radial vector is perpendicular to the contour. Figure 5.1 illustrates these definitions.

The internal force is based on the local shape of the contour, and aims at minimizing local curvature. The presence of the internal force will force the contour to keep a more or less circular shape. Local curvature is computed for vertex i using the two adjacent edges computed by subtracting the unit vectors $\hat{d}_i$ and $\hat{d}_{i-1}$

$$c_i = \hat{d}_i - \hat{d}_{i-1}.$$

Local curvature therefore has both a strength (the length of the vector) and a direction equal to the radial vector $\hat{r}_i$ or the opposite direction. The curvature depends on the angle between the two edges, not on their length. To prevent the contour from imploding, vertices in parts

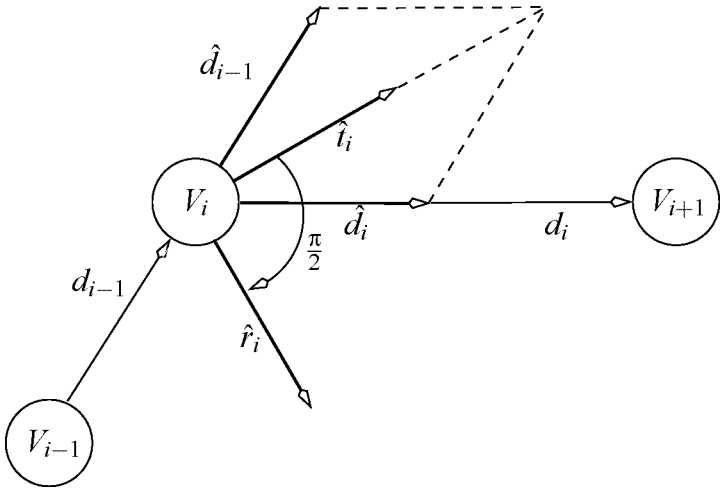

Figure 5.1: Example showing a part of a contour. The radial vector $\hat{r}_i$ and the tangential vector $\hat{t}_i$ for vertex V_i are shown. Note that $\hat{d}_i$ denotes the unit vector of d_i.

of the contour with constant curvature should have an internal force of zero. To achieve this, the internal force of a vertex is computed by combining its local curvature with the local curvature of the two neighbor vertices in the local coordinate system:

$$f_{int,i} = (-\frac{1}{2}(c_{i-1}\hat{r}_{i-1}) + c_i\hat{r}_i - \frac{1}{2}(c_{i+1}\hat{r}_{i+1}))\hat{r}_i.$$

For example, for the polygon in Figure 5.2, all vertices have different curvature vectors in the Cartesian coordinate system but they are the same in the $(\hat{r},\hat{t})$ coordinate system, giving an internal force of 0 for every vertex.

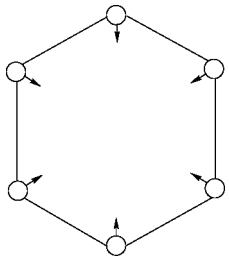

Figure 5.2: A contour with 6 vertices with their curvature vector. All curvature vectors are equal in the (r,t) coordinate system.

The external force is determined by the image gradient magnitude. This force will move the vertices to locations in the image with strong gradients: the edge of the tumor. For the pixel where vertex i is located, the second order gradient derivative is computed. Computing the external force is done in the radial direction, because this will prevent vertices from moving along the contour. An example of the discrete contour model applied to a mass is shown in Figure 5.3. The paths of the individual vertices are shown in the middle figure, and the final contour is shown in the right curve. The contour that is found for this case is very close to the annotation.

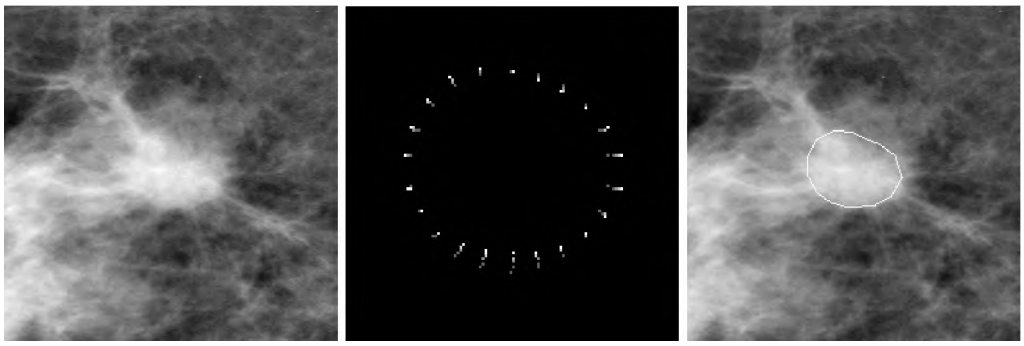

Figure 5.3: The left figure shows a clearly visible mass. When the discrete dynamic contour model is applied the vertices follow the paths that are shown in the middle figure (enlarged), resulting in the contour in the right figure.

The success of the discrete dynamic contour model depends strongly on the initial contour. Because masses are more or less circular, we initialize the method with a circular contour with a fixed size appropriate for masses. A number of parameters has to be set for the discrete contour model. The balance between the internal and external force can be set with weight parameters. A preference for the internal force will give smooth contours, a high value for the external force weight parameter will yield more capricious boundaries. Another important parameter is the scale at which the gradient is computed. A large value for the scale parameter results in a more robust method but limits the accuracy. On the other hand, when the scale is chosen small, the initial contour must be located accurately, otherwise the method will fail to converge to the edge of the mass.

5.3 Region growing

Region growing is one of the most popular methods to segment regions in images [16]. The basic idea of region growing is very simple. Given a starting pixel or region, connecting pixels or regions are added if their properties are similar to the already segmented region. In most applications the chosen property is simply the intensity value.

Assume a region of interest I with intensity values described by $f(x,y)$. A seed point $(x_s,y_s) \in I$ with an intensity value of $f(x_s,y_s)$ is used as the starting point and is the initial segmented area. All neighboring pixels with an intensity value larger than a threshold are added to this segmented region. This procedure is repeated until no more pixels are found with values above the threshold. A contour is found and the whole procedure is repeated with a lower threshold value, resulting in a larger segmented region. This way, a series of contours is created.

Several criteria can be used to select the best contour, for example based on shape measures. Another simple option is to use a fixed threshold level to stop the growing process. In this work, two different criteria were implemented: a fixed threshold value that depends on $f(x_s,y_s)$ and a probabilistic method developed by Kupinski and Giger [4]. Based on estimations of the probability distributions of intensity values of background tissue and pixels inside a tumor, for each contour that is found in the growing process a likelihood is computed. The contour with the maximum likelihood value is selected as the best contour.

One of the main problems with region growing in this application is that the prior knowl-

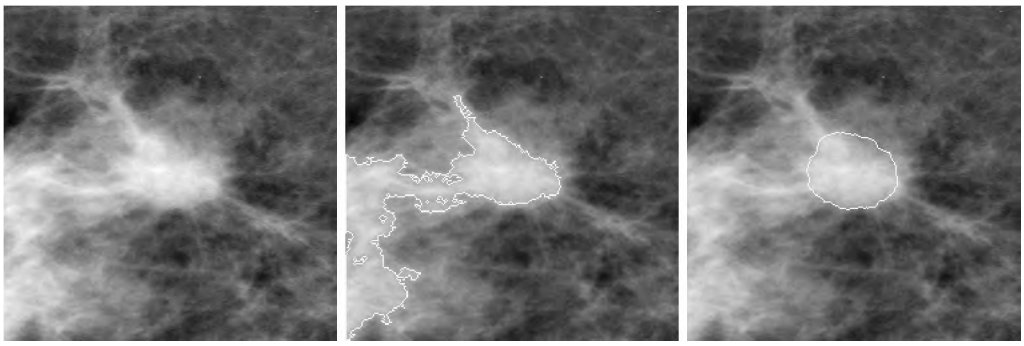

Figure 5.4: The left figure shows the same mass as in Figure 5.3. When region growing is applied without preprocessing, the contour in the middle figure was obtained. When the region was preprocessed, the contour in the right figure was found.

edge that the segmented regions should be more or less circular is not incorporated in the growing process. To include this knowledge, Kupinski and Giger first preprocessed the region. A Gaussian function, centered at (x_s, y_s), was used in a multiplication, giving each pixel (x, y) in I the value

$$h(x,y) = f(x,y)N(x,y,x_s,y_s,\sigma^2).$$

Distant pixels are now suppressed, resulting in more circular regions. Also, a good value for σ will yield regions with sizes that are common for masses. The value of σ is important, especially when fixed thresholds are used to segment the region. A value of 14 mm, similar to the value used by Kupinski and Giger, was found to give best performance. Figure 5.4 shows an example of the region growing procedure, both with and without preprocessing.

5.4 Experiments

Both mass segmentation methods described in the previous section need a starting point. First, the centers of gravity of the annotations was used for this purpose. Generally, these points will be close to the optimal starting points and a good comparison between the methods can be made. In a second experiment, a pixel-based mass detection method was used to generate starting points [14]. The used method assigns a measure of suspiciousness to each pixel in the mammogram. For each mass, the location inside the annotation with the highest measure of suspiciousness was used as the starting point for the segmentation methods. These points may be located less central than the starting points in the first experiment. The robustness of the methods to these suboptimal starting points was examined. It should be noted that the performance of the methods in this second experiment depends strongly on the algorithm used in the first step.

To evaluate the performance of the segmentation methods the following overlap criterion was used

$$\text{Overlap} = \frac{S \cap T}{S \cup T},$$

where S is the segmented area and T is the annotation made by the radiologist. A value close to one means a good match between the two regions. To visualize the performance of

the methods, the results will be presented in the same way as was done Kupinski and Giger, where the success rate is shown as a function of the overlap. Horizontally the overlap is shown, vertically the fraction of tumors for which the method achieved at least this overlap. Robust methods will have a reasonable overlap in a high percentage of cases. For only a few cases, the method will fail to find a reasonable segmentation. Accuracy of the methods can be judged by the percentage of cases that have a large overlap, for example more than 0.7.

5.4.1 The data set

A set of 136 women with a total of 214 mammograms was used to test the performance of the segmentation methods. This set is a combination of a number of sets of different origin including spiculated, circumscribed and vague masses. The majority, 194 mammograms, were found in the Dutch screening program, 20 mammograms with a malignant mass were taken form the MIAS data set [12].

The recording and digitalization of the mammograms varied due to their different source and time of acquisition, but all were reduced to 200 μm per pixel. All tumors were classified and annotated with the help of an experienced radiologist. Only the central area was annotated, spicules were left out.

5.4.2 Centers of gravity as starting points

In this experiment the center of gravity of the annotation made by the radiologist was used as the starting point. The region growing process was done both with and without the preprocessing step. The basic version used a threshold that was a percentage of the intensity value of the pixel at the seed location. Without preprocessing, the best results were obtained with a threshold at 94%.

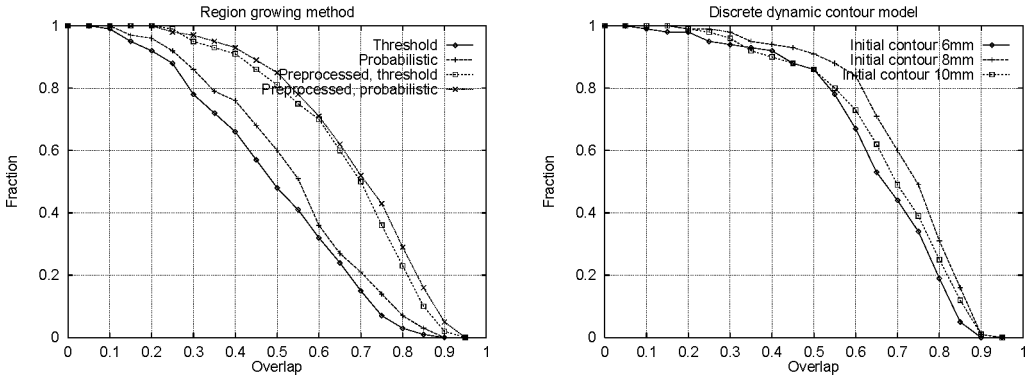

Figure 5.5: Results for the methods when for each mass the center of gravity is used as a starting point. The left figure shows the results of the region growing method, the right figure the results for the discrete dynamic contour model.

In Figure 5.5a, the results for the region growing method are shown. After preprocessing, a lower threshold should be used, otherwise very small regions will be the result. Preprocessing with σ=14 mm and a threshold of 80% gave best performance, which was

considerably better than the performance that was achieved without preprocessing. The probabilistic method for contour selection performed better than the simple thresholding method in the situation without preprocessing. After preprocessing, a more or less similar performance was achieved.

The discrete dynamic contour model used a fixed sized circle as its initial contour. Three curves for various radii are shown in Figure 5.5b. In Figure 5.6a the best curves for both methods are compared. The discrete dynamic contour model outperforms the region growing method. Figure 5.6b shows a scatter plot where each point represents a mass from the data set for the two curves in Figure 5.6a. Horizontally the obtained overlap for the discrete contour model is shown, vertically that for the region growing method.

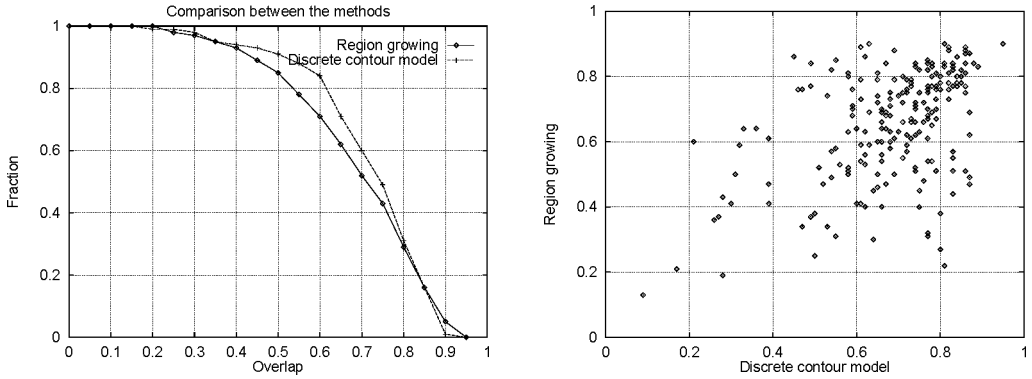

Figure 5.6: The best curves for both methods with the center of gravity of the annotations as starting point. The discrete contour model performs better than the region growing method. This is also visible in the asymmetry of the scatter plot.

5.4.3 Automatically generated starting points

When mass segmentation is used in a false positive removal procedure, normally it is in a second step after an initial detection method. An initial pixel-based detection step [13, 14] was used to select suspicious areas, assigning each pixel a measure of suspiciousness. For each mass, the pixel inside the annotation with the highest measure was selected as the starting point. Rather different results are obtained in this experiment compared to the previous experiment, as is shown in Figure 5.7 and Figure 5.8. Less gain is achieved for the region growing methods when the region is preprocessed by the multiplication with the Gaussian. Although preprocessing is still beneficial, the gain is much lower. The probabilistic contour selection method performs better than the simple threshold when the region is not preprocessed, with preprocessing the performance is slightly lower. Figure 5.8 shows that the discrete dynamic contour model outperforms the region growing method in this experiment.

5.5 Discussion

Region growing benefited much from the preprocessing step with the Gaussian. When the center of the annotation was used as the starting point, the performance with preprocessing was higher than without. Probabilistic contour selection performed better than the threshold version when the image was not preprocessed. With preprocessing, a similar performance

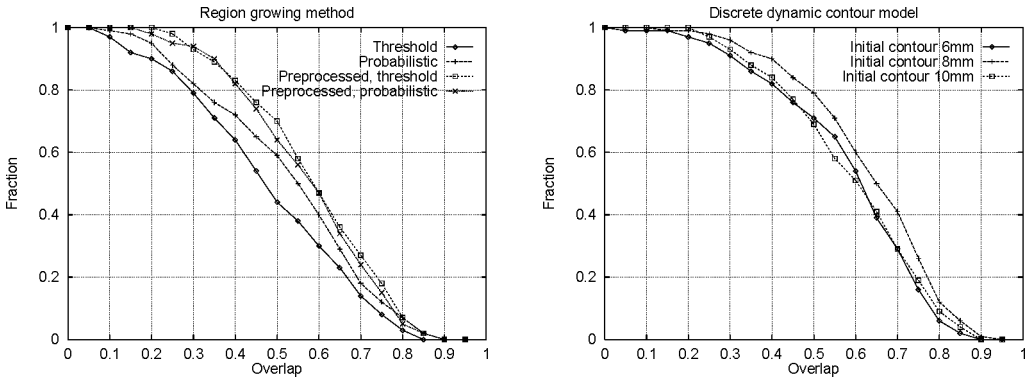

Figure 5.7: Results for the methods when for each mass the output of the pixel-based mass detection method is used as a starting point. The left figure shows the results of the region growing method, the right figure the results for the discrete dynamic contour model.

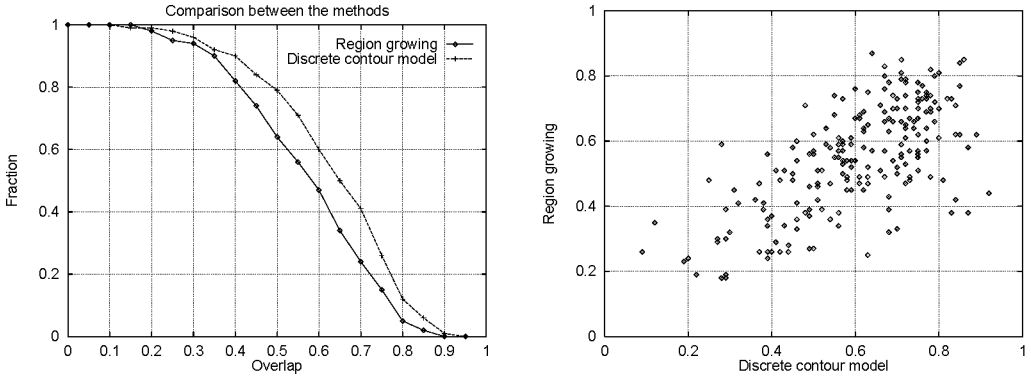

Figure 5.8: The best curves for both methods compared when the output of the pixel-based mass detection method is used as a starting point. The discrete contour model performs better than the region growing method. Again, the scatter plot is rather asymmetric.

was achieved. When the output of the pixel-based detection method was used to generate starting points, the preprocessing step was much less beneficial. In many cases the center of the Gaussian was located suboptimal which makes accurate segmentation of the mass more difficult. The region growing methods without preprocessing had similar performance as in the first experiment. The same contour was found for almost every case. Preprocessing still increased the performance, but the gain was smaller than in the first experiment. Probabilistic contour selection performed similar to the thresholded version when the region was preprocessed, probably due to a suboptimal estimation of the mean of the tumor-pixel distribution.

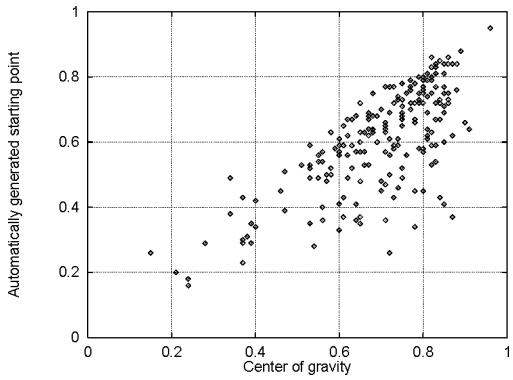

Figure 5.9: Scatter plot of the performance for each mass in both experiments for the discrete dynamic contour model. Horizontally the overlap in the first experiment, vertically the overlap of the second experiment is shown. A strong asymmetry is visible: for many tumors a worse segmentation is found in the second experiment.

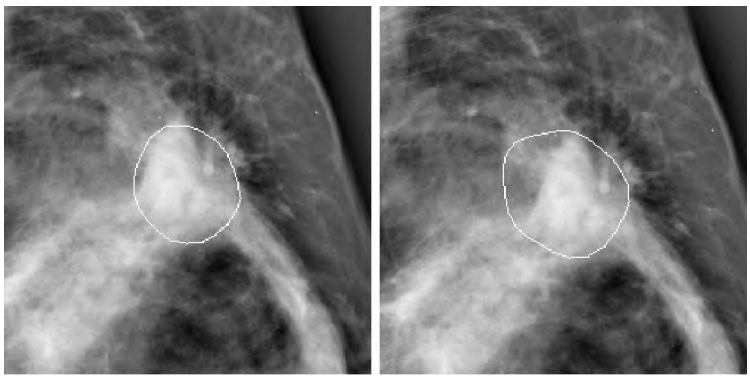

Figure 5.10: For the contour in the left figure the center of gravity was used, for the contour in the right figure the generated peak. In the upper left part of the tumor, the method fails to converge to the edge.

The discrete contour model performed better than the region growing method in both experiments. However, its performance decreased considerably when the starting point was not chosen at the center of the annotation but automatically generated. In Figure 5.9 the performance in both experiments is shown for all 214 masses. An example were a worse performance was found when the peak was used instead of the center of gravity is shown in Figure 5.10. For this mass, the performance decreased from 0.82 to 0.59.

Both methods rarely achieve more than 85% overlap. The first reason for this is that the annotations are on the large side to make sure the whole tumor area is inside the annotation. The second reasons is that when the annotation and the segmented area are not identical, the chosen overlap criterion quickly decreases.

5.6 Conclusions

The discrete contour model is a robust method to segment masses, and outperformed a probabilistic region growing method. Especially when the initial pixel-based mass detection

method was used to generate the starting points the discrete contour model performed better. The experiments show that it is a promising approach to use in a false positive removal procedure. However, just like for the region growing methods a good choice for the seed point is important, the success of the segmentation depends strongly on the accuracy of the initial pixel-based step.

Acknowledgments

This project was supported by the Dutch Cancer Society (Koningin Wilhelmina Fonds) under Grant KUN 96-1343.

Bibliography

[1] M L Comer, S Liu, and E J Delp. Statistical segmentation of mammograms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 475–478. Elsevier, Amsterdam, 1996.

[2] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[3] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[4] M A Kupinski and M L Giger. Automated seeded lesions segmentation on digital mammograms. *IEEE transactions on medical imaging*, 17(4):510–517, 1998.

[5] H D Li, M Kallergi, L P Clarke, V K Jain, and R A Clark. Markov random field for tumor detection in digital mammography. *IEEE Trans on Med Imag*, 14:565–576, 1995.

[6] S Lobregt and M A Viergever. A discrete dynamic contour model. *IEEE transactions on medical imaging*, 14(1):12–24, march 1995.

[7] T McInerney and D Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2):91–108, 1996.

[8] A J Mendez, P G Tahoces, M J Lado, M Souto, and J J Vidal. Computer-aided diagnosis: automatic detection of malignant masses in digitized mammograms. *Medical Physics*, 25(6):957–964, June 1998.

[9] N Petrick, H P Chan, B Sahiner, D Wei abd M A Helvie, M M Goodsitt, and D A Adler. Automated detection of breast masses on digital mammograms using adaptive density-weighted contrast enhancement filtering. *SPIE 2434*, pages 590–597, 1995.

[10] S Pohlman, K A Powell, N A Obuchowski, W A Chilcote, and S Grundfest-Broniatowski. Quantitative classification of breast tumors in digitized mammograms. *Med Phys*, 23:1337–1345, 1996.

[11] M Sameti and R K Ward. fuzzy segmentation algorithm for mammogram partitioning. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 471–474. Elsevier, Amsterdam, 1996.

[12] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R

Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[13] G M te Brake and N Karssemeijer. Detection of stellate breast abnormalities. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 341–346. Elsevier, Amsterdam, 1996.

[14] G M te Brake and N Karssemeijer. Automated detection of breast carcinomas not detected in a screening program. *Radiology*, 207:465–471, 1998.

[15] F F Yin, M L Giger, C J Vyborny, K Doi, and R A Schmidt. Comparison of bilateral-substraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol*, 6:473–481, 1993.

[16] T Y Young and K S FU (eds). *Handbook of pattern recognition and image processing*. Academic Press Inc., London, UK, 1986.

Chapter 6

Specificity improvement by regional analysis for mass detection algorithms[1]

6.1 Introduction

False negatives in mammography are often related to masses and architectural distortions that are overlooked or misinterpreted [2, 17, 4]. Masses and distortions can be extremely subtle and are often covered by normal tissue due to superposition, sometimes just a faint density or a slight distortion of the tissue is visible. Because of the importance of detecting tumors at an early stage of development, many research groups are developing algorithms for mass detection to aid radiologist with this problem. A variety of approaches has been suggested, but most follow the two-step scheme of pixel-level detection and region-level classification as described by Woods and Bowyer [19]. First, a pixel-level detection method is used to detect suspicious areas in mammograms. In a region-level detection step these areas are examined more closely and classified normal or abnormal. Some groups put much intelligence in the first step and continue with a very basic classification step, while other groups use very simple techniques like thresholding to detect suspicious regions and use a complex classification step to remove false positive signals.

The focus of our work so far has been towards the first pixel-level detection step. Our approach to detect masses aims at detecting both spicules and the central mass [7, 15]. Presence of either one or both of these properties triggers the system to signal a suspicious region. This detection algorithm is very sensitive, of a set of masses that occurred in the Dutch screening program over 98% were found at a false positive level of approximately 4 FP/image. However, clinical application of detection software requires higher specificity levels. Aim of this work is to improve the performance of the detection method by applying a region-level classification step to remove false positive detections.

A number of groups have been working on the problem of classifying suspicious structures in normal or abnormal types. Most groups apply algorithms to segment the suspicious region [7, 13, 12, 9, 20], but sometimes texture features are computed over a large region containing the suspicious region [18]. Segmentation of the suspicious region is useful in separating abnormal and normal tissue, because it enables computation of features related

[1] G.M. te Brake, N. Karssemeijer, J.H.C.L. Hendriks, *An automated method to discriminate malignant masses from normal tissue in digital mammograms*. Submitted to Physics in Medicine and Biology.

to the edge of the region, as well as contrast and shape features. In this work, a discrete dynamic contour model was used to segment the regions, which has proven to be a robust and fast method for this task [15].

Which features will be successful in a false positive rejection step depends on the types of regions that are selected by the initial detection step. For instance, if the detection step generates many false positives on crossing lines, a feature that detects ducts may be useful. On the other hand, if all bright areas are signaled, shape analysis of these regions may be a useful approach. Rejecting false positives is not the same as classifying between benign and malignant masses, for which edge analysis is very important. Removal of false positives may require other types of features because many are due to projection or are induced by the edge of the pectoral muscle.

The features that are used in this work to separate normal from abnormal tissue are related to image characteristics that are used by radiologists. Instead of using complex texture measures that may depend on the film or digitizer that was used, our features are related to properties of the suspicious region that the radiologists incorporate in their decision process as well. It is likely that these features are also successful after other initial detection methods and are relatively independent of the way the images were acquired.

An overview of the system is presented in Figure 6.1. The initial pixel-level detection step creates a likelihood image for the mammogram, an image where each pixel is assigned a measure of suspiciousness. A peak detection method is applied to this likelihood image. For each peak, a contour is fitted to the suspect region, and for each segmented area a number of features are computed. Finally, a neural network uses these features to compute a final measure of suspiciousness, after which the performance of the method can be shown by Free response operating characteristic (FROC) curves.

In the next section, a short description is given of the initial detection step and the method that is used to segment the suspicious area. Section 6.3 describes the features that are computed using the contour. Section 6.4 describes a number of features that do not depend on the contour, like the distance of the peak to the skin line, or that are related to the number of high peaks in the mammogram. These features will be referred to as peak-related features. The experiments and the data sets that are used are described in Section 6.5, the results are presented in Section 6.6.

6.2 Detection and segmentation

To classify suspicious regions in mammograms, they must be detected and segmented. The methods that are used in this work to detect and segment the regions have already been published, and therefore are only described briefly in the next two subsections.

Before the detection algorithms are applied to the image, the breast area is segmented automatically. The skin line is found, as well as the edge of the pectoral. Near the skin line the breast is less thick, which causes a fall-off in intensity which is corrected for. Also the projection of the pectoral muscle is subtracted from the image when present, a method that was described in [8].

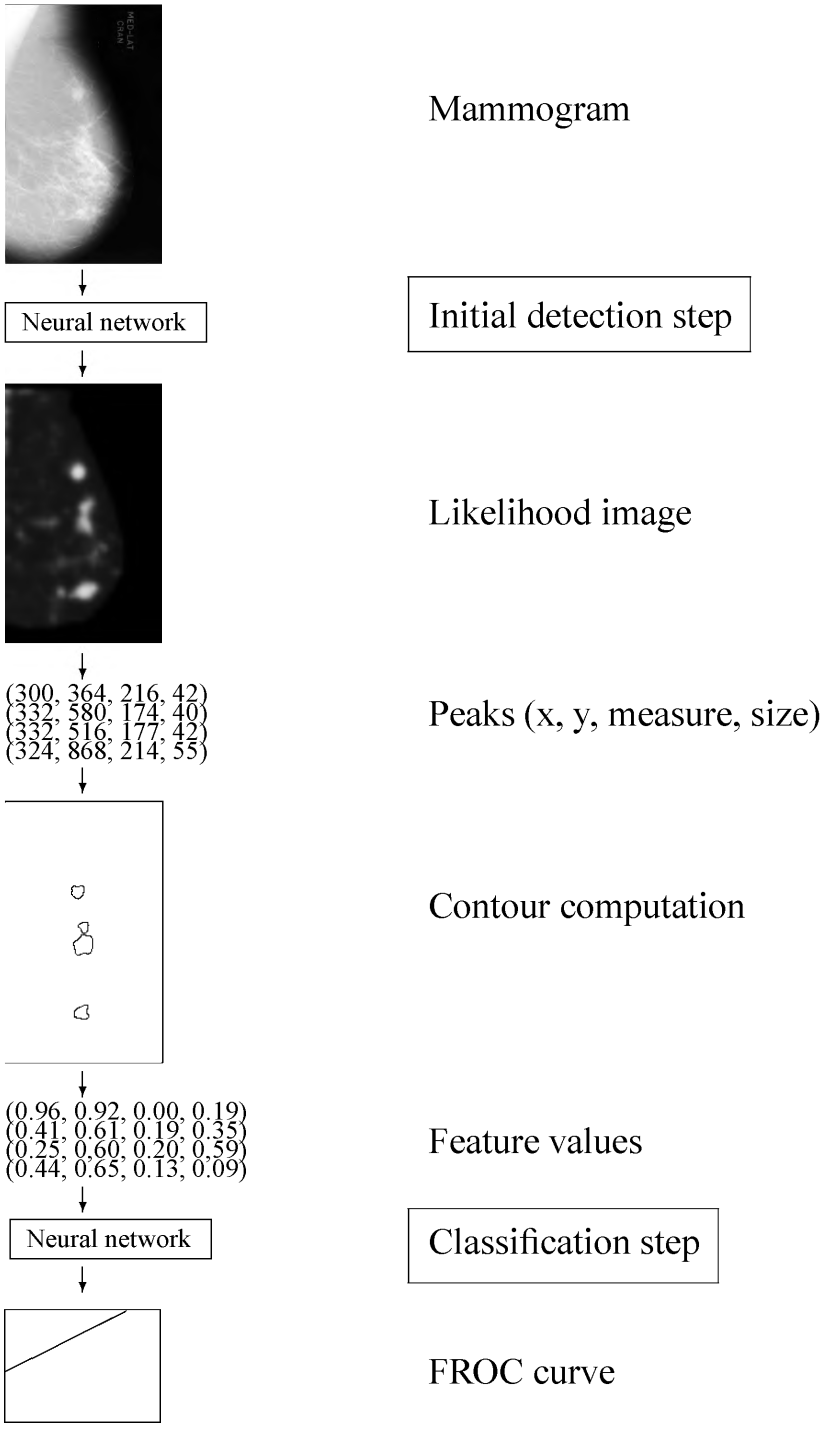

Figure 6.1: Overview of the system.

The initial pixel-level mass detection step

The method that was used to detect suspicious regions was developed in our group [7, 15]. It is a pixel-level method: each pixel is assigned a measure of suspiciousness for malignancy. Two image characteristics are very important for tumor detection: the presence of a central mass, and the presence of a radiating pattern of spicules. Both characteristics are detected with our method, making it sensitive even to small and subtle tumors. Five features were computed for each pixel, three related to the presence of spicules, two related to detecting the central mass. A neural network classified each pixel using these 5 features and assigned a measure of suspiciousness to it, yielding a likelihood image. This likelihood image was slightly smoothed, and the highest peaks in this likelihood image were selected. For each peak that was found, its location was saved together with its measure of suspiciousness and a size estimate of the found density. This size estimate was determined by examining the scale at which the mass features gave the highest value.

Mass segmentation

In recent years, deformable models have become popular in the field of medical image analysis [11]. We have applied a member of this family, a discrete dynamic contour model, to the task of mass segmentation. The implementation of the contour model was based on an algorithm described by Lobregt and Viergever [10], a fast and robust procedure to detect boundaries. It is a discrete model represented by vertices that are connected by edges. An initial contour has to be chosen, after which each vertex is moved around by a combination of internal and external forces working on it. These forces determine the speed and acceleration of the vertex. Previous research has shown that this approach was superior to region growing for the task of mass segmentation [16], and therefore this method was also used in this work.

For each peak that was found by the initial detection step, an initial contour was generated using the size estimate of the tumor. This was an improvement of the previously described method, where a fixed initial circle size was used. Another improvement was that the gradients that were used to compute the external force were computed at two different scales: first at a large scale to move the contour to the edge of the tumor, followed by a smaller scale for a more accurate fit.

6.3 Contour-related features

Radiologists use a number of image characteristics to determine whether a suspicious looking region is normal or requires further examination. Important characteristics are:

Intensity and contrast If the region has high contrast or a higher intensity than other similar structures in the image it is likely to be a mass.

Isodense Tumors are more or less isodense objects, and cannot be seen trough. If a regions has holes, it is likely to look suspicious due to unfortunate projection of normal tissue.

Location If the region is located in a fatty surroundings, this is more suspicious than when it is part of the normal glandular tissue area. Areas like the lower medial area are supposed to contain fat, dense tissue is suspicious.

Texture A pattern of lines radiating around the region is an important sign of malignancy. If these lines are going trough the center area, they are more likely present due to the projection of normal tissue or ducts and make the region less suspicious.

Deformation of the skin line or of the glandular tissue Malignant abnormalities deform the normal structures in the breast, causing retraction of the skin or deformations in the glandular tissue.

Appearance in both oblique and cranio-caudal view If a density is only visible in one view, it may be caused by superposition of normal tissue. On the other hand, if a density is visible in both views it is likely to be a real lesion.

Asymmetry Asymmetry between the left and right mammograms can indicate abnormal tissue.

Temporal changes When structures are found in mammograms that were not present on mammograms taken two years before, this is suspect because mammograms should become fattier over time. Exceptions to this rule are women who use hormone replacement therapy.

In this work, features are defined and tested that aim at capturing the first four characteristics. Detection of deformations of the skin line or the glandular tissue is a complicated problem, due to high variations in appearance and the unspecific nature of this sign, and was not part of this work. No work has been done yet in our group on correlating regions that are found in two views of the same breast. For the last two characteristics, two mammograms must be compared. Comparing mammograms is a complicated task, because due to differences in compression two mammograms from the same breast can look rather different. Some initial work has been done in our group on this topic [8], but this has not been integrated in this work which focuses on computing features based on the found segmentations.

Features based on image characteristics that radiologists use are likely to be relevant for other initial detection steps as well and are intuitively understandable by radiologists. They are probably more robust to variations in the acquisition process than many of the texture features that are often used for this type of tasks.

To compute features based on the contour such as contrast, differences between the area inside and outside the contour must be determined. For this purpose, an area outside the contour was defined. The effective radius of the segmented region was approximated by

$$R = \sqrt{\frac{\text{Size of the segmented area}}{\pi}}.$$

All pixels within a distance of $0.6R$ to the segmented area formed the outside region that was needed to compute the features. The outside area is approximately twice the size as the segmented area, an example is shown in Figure 6.2. The segmented area is white, the grey pixels form the outside area. In this section, the set of pixels in the segmentation is denoted by I, the set of pixels on the contour by C and the pixels in the outside region by O. $E(X)$ is the mean value of pixels in set X, $Var(X)$ their variance, and $N(X)$ denotes the number of pixels in set X. Finally $H(X,i)$ denotes the fraction of pixels in set X with intensity value i. To limit the number of entries, the intensity range was divided into 82 bins, each containing a range of 50 grey values.

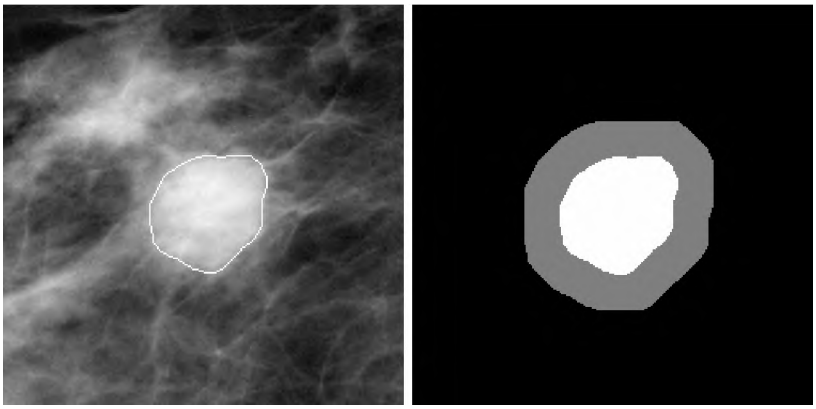

Figure 6.2: A malignant mass with the final contour. The right image shows the segmented and outside area that were used to compute the features.

Intensity and contrast

Tumor tissue absorbs much more x-rays than fat and slightly more than glandular tissue, and therefore intensity and contrast are expected to be useful features for removing false positive signals. Many different contrast measures are described in the literature, of which 6 commonly used measures were selected for this work. All mammograms were digitized with the same type of digitizer (Lumisys model 85 or Lumisys model 200), using a fixed linear relation between pixel values and optical density. Therefore, no scaling was required to bring all mammograms in the same intensity range, and simple contrast measures can be used.

The first feature that was computed was the mean intensity in the segmented area:

$$\text{Intensity} = E(I).$$

The most simple contrast measure that we used is the difference in intensity between I and O;

$$\text{Contrast1} = E(I) - E(O).$$

The second contrast feature is a distance measure between the two grey-value histograms of I and O. The square of the difference between both means was divided by the sum of the variances of the two areas,

$$\text{Contrast2} = \frac{(E(I) - E(O))^2}{Var(I) + Var(O)}.$$

The third contrast value also represents the distance between the two histograms. For each entry in the normalized histogram of the two areas, the absolute value of the difference was computed:

$$\text{Contrast3} = \sum_i |H(I,i) - H(O,i)|.$$

This yields a value between 0 (total overlap) and 2 (complete separation). The main advantage of this feature is that it is independent of scaling of the intensity values of the image.

Three other well-known difference measures were computed on the two histograms, divergence:

$$\text{Contrast4} = \sum_i (H(I,i) - H(O,i)) \ln \frac{H(I,i)}{H(O,i)},$$

the Bhattacharyya coefficient:

$$\text{Contrast5} = -\ln \sum_i (H(I,i)H(O,i))^{-\frac{1}{2}},$$

and the Matsutsita distance

$$\text{Contrast6} = \sqrt{\sum_i (H(I,i) - H(O,i))^2}.$$

Practical problems arise when the divergence measure and the Bhattacharyya coefficient are computed because most bins in the two histograms are empty. To solve this, only bins i for which both $H(I,i)$ and $H(O,i)$ were non-empty were used in the computation.

Isodense

Suspicious looking regions can sometimes be classified as normal tissue when holes are present inside the area. An example is shown in Figure 6.3. Holes indicate that the region is suspicious due to projection, because tumors are normally dense and cannot be seen through. An exception is the lobular carcinoma that can cause architectural distortion of the tissue in the breast. To determine whether dark areas are present in the segmented area, a feature was developed that examines whether the low values in the segmented area are higher than the pixel values found in the surrounding tissue. If this is not the case, it is possible to look through the segmented area, suggesting a false positive region due to projection. This features relies on an accurate segmentation, because if part of the surrounding is segmented as well, the feature yields incorrect results.

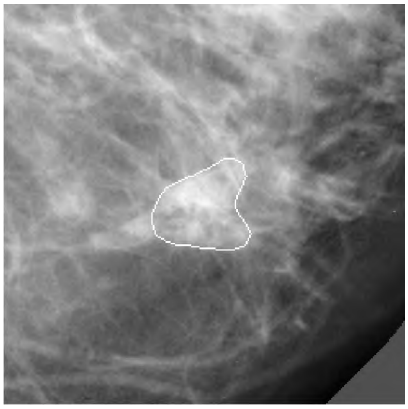

Figure 6.3: A normal structure that looks suspicious due to projection, but that is not isodense and therefore unlikely to be a tumor.

The feature that is computed is the fraction of pixels in area O that has a lower value than a threshold t. The threshold is determined by the intensity level for which 10% of the pixels inside the segmented area is lower. When the fraction of pixels is close to 1, it indicates the presence of a tumor. More precisely, the largest t is found for which

$$\sum_{j=0}^{t} H(I,j) < 0.1,$$

yielding the feature value

$$\text{Isodense1} = \sum_{j=0}^{t} H(O,j).$$

The inverse of this feature was also computed: the fraction of pixels inside the segmented area that is larger than the 90% threshold on the intensity values from pixels in the outside area. Again, the largest value t for which

$$\sum_{j=0}^{t} H(O,j) < 0.9$$

is found, yielding the second isodense feature value

$$\text{Isodense2} = 1 - \sum_{j=0}^{t} H(I,j).$$

These features do not incorporate the topology of the lowest 10% of the pixels. They can be near the edge of the segmentation, or located in a number of holes in the area. It is possible that more advanced topology based methods better reflect the concept of isodenseness.

Location

A density is suspect if it is found in a fatty area separated from the rest of the dense tissue present in the mammogram. Even if it is small and it has low contrast, it is suspicious because these areas should be free of dense tissue. An example of such a tumor is shown in Figure 6.4. This tumor is visible due to its location, tumors of this size and contrast that are embedded in the glandular tissue will not be detected.

A method to segment the dense tissue region in the breast has been developed, and was used for this purpose. It is a simplification of the approach described by Aylward [1]. A mixture model of two Gaussians is fitted to the histogram of grey values of the mammogram, one Gaussian representing fat, the other representing dense tissue. Pixels for which the likelihood of belonging to the dense region is high are segmented. An example is given in Figure 6.5, a mammogram from the Nijmegen set with a low contrast tumor visible in fatty surroundings. The fraction of pixels within a distance of $2R$ of the segmented area that was inside this normal tissue area was computed, with R the effective radius of the segmented region. Low values indicate that the signal is inside a fatty area, and therefore more suspicious. The same feature was also computer for a larger region, including all pixels with a radius of $4R$.

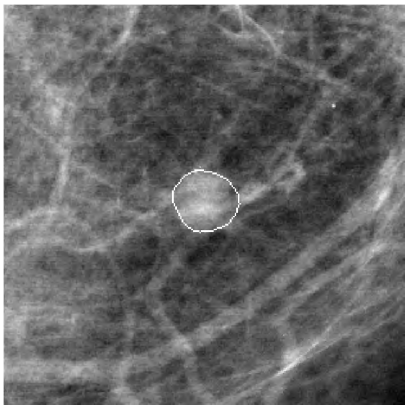

Figure 6.4: A small tumor that is visible due to its location in a fatty area.

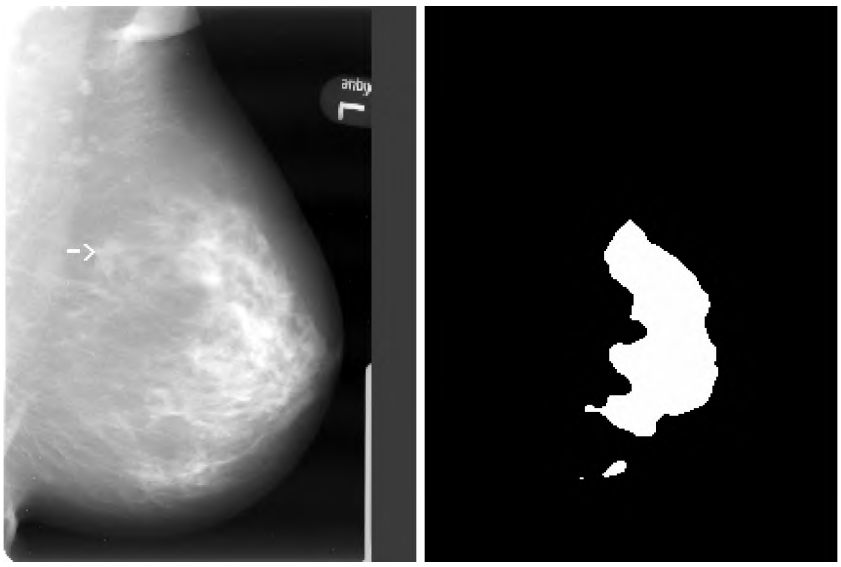

Figure 6.5: A mammogram from the Nijmegen set with the segmentation of its normal glandular tissue. A tumor is present at the arrow.

Linear texture

If a region is suspicious due to projection, linear structures are often found inside the segmented area. An example of a normal structure found in a mammogram with this property is shown in Figure 6.6. A texture measure was computed to capture the linear structure in a segmented region. The feature is computed by making an estimate of line orientation for each pixel inside the segmented area. This estimate is a vector representing line orientation and line contrast. If no line is present the magnitude will be low and the orientation random. All vectors are summed using the double angles representation [5], giving a final "total"-vector. The first feature is the length of this total-vector, the second feature is this length divided by the total length of all vectors. The lines estimate was computed at 2 different scales using second order derivatives of the Gaussian function, with a sigma of 1 pixels and a sigma of 3 pixels.

An overview of all features described above is given in Table 6.1.

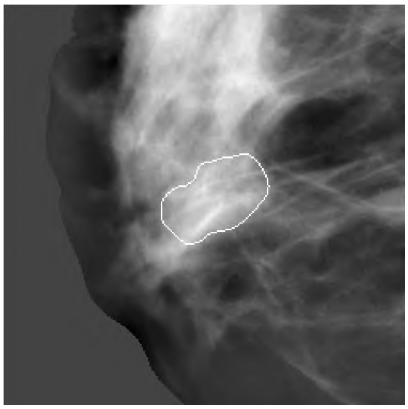

Figure 6.6: Projection with much linear structure inside segmented area.

6.4 Peak-related features

To classify the regions, a number of features were used that did not depend on the contour. The first feature was the likelihood measure that was produced by the initial mass detection method. This likelihood value include spicule information that is required in the second classification step.

When a high peak is found in the likelihood image and no other high peaks are present, it is more likely to signal a malignant mass, than when many other high peaks are present. In the latter case, high peaks are more likely due to properties of the glandular tissue. Therefore, a normalized version of the likelihood feature was computed for each peak based on the other detections in the same mammogram. The average likelihood measure of peak 5 to peak 8 in the mammogram was computed. For each peak, its likelihood measure was divided by this average value to compute the new feature. If a peak has a measure that is much higher than that of other peaks, its normalized likelihood measure will be high. The first 4 peaks were not incorporated into this normalization because when a tumor is present it is almost always among these peaks. In this way, it is avoided that the presence of the tumor influences the normalization of the normal regions in the same mammogram, which would give a positive bias to the computed specificity.

A third feature that was generated was the distance from the detected peak to the skin line. In the initial detection step, false positives are frequently found close to the skin line, a location were only few tumors occur. It is hoped that this feature will help to remove a number of these false detections. However, care must be taken to prevent that tumors located just behind the nipple (a location were tumors are often found) are suppressed by this feature.

The 3 peak-related features are presented in Table 6.2.

Data sets

Two data sets were used in this work to test the region-level classification method. First, a consecutive set of 71 cases taken from the Dutch screening program was constructed. All cancers were included, except those cases where a cluster of microcalcifications was the

Name	Description
Intensity	Mean pixel value inside contour
Contrast1	Difference in mean value inside and outside contour
Contrast2	Normalized contrast measure
Contrast3	Distance contrast measure
Contrast4	Divergence measure
Contrast5	Bhattacharyya coefficient
Contrast6	Matusita distance
Isodense1	Isodenseness of the segmented area
Isodense2	Inverse isodense measure
Location1	In fatty or dense area, small area size
Location2	In fatty or dense area, large area size
LinearTexture1	Presence of linear texture
LinearTexture2	Presence of linear texture, normalized
LinearTexture3	Presence of linear texture, large scale
LinearTexture4	Presence of linear texture, large scale, normalized

Table 6.1: Overview of the contour-related features

Name	Description
Likelihood	Measure of suspiciousness of first detection step
NormLikelihood	Normalized version of the first feature
Distance	Distance of the peak to the skin line

Table 6.2: Overview of the peak-related features.

only visible sign. All women participating in this program are between 50 and 69 years old. If a woman is screened for the first time both oblique and cranio-caudal films are made. On succeeding visits, cranio-caudal films are only made when the radiographers find the oblique films hard to read due to dense tissue, or when they find a suspicious area. For each case in the database the oblique films were also included, in 61 cases cranio-caudal films were also available. This makes a total of 132 mammograms with a visible malignant tumor, ranging from very subtle to very obvious, but a typical sample of tumors that occur in screening. The contra-lateral films were included, as well as 208 normal mammograms for a better estimate of the specificity. This makes a total of 472 films, with 132 malignant abnormalities.

The second set that was used was a part of the new Digital Database for Screening Mammography (DDSM) [6, 3]. All the malignant masses present in the sets "cancer_01", "cancer_02"and "cancer_05"were used, including their contralateral images. Again, cases with only microcalcifications were left out, making a total of 193 cases: 386 oblique films and 386 cranio-caudal films. For this set, 372 annotations of malignant lesions were given, in 400 films no malignancies were found. In a few cases, the lesion was visible in only one view. In these cases the other view is included in the set as a normal film. The annotations of the cancers that come with the DDSM database are extremely large. To prevent that signals due to normal tissue are inside the annotated area and are counted as detections, tighter annotations were made.

Our principal interest is detection of malignant abnormalities. Benign abnormalities were not annotated, and therefore will induce a number of false positive signals.

6.5 Design of the Experiment

The initial pixel-level detection method was trained on a data set containing 39 mammograms taken from the MIAS dataset [14], and applied to the two data sets that were used to test the region-level classification method. On average, 5 false positive areas were detected per image (ranging from 0 and 20), a specificity level at which most tumors are detected. For each detected area, the size estimate of the suspect region was used to estimate an appropriately sized initial contour, and the region was segmented using the dynamic contour model.

The next subsection describes how the regions were classified and the how the performance of the methods was evaluated. The feature selection method that was used to find the best feature for each image characteristic is described.

Classification

Neural networks were used to classify the regions based on the computed peak-related and contour-related features. Simple 3-layer feed-forward neural networks trained using the back-propagation algorithm were used for this purpose. The number of input units was equal to the number of features. In all experiments 3 hidden nodes and two output nodes (one for normal, one for abnormal regions) were used. The difference of the two output units was used as the measure of suspiciousness. To minimize variations due to different training runs, for each feature combination five neural networks were trained and their output was averaged. When the DDSM set was used for testing, the Nijmegen set was used to train the classifiers, and vice versa.

The average value of the output of the 5 neural nets was thresholded at various levels, which yielded free response operating characteristic (FROC) curves for the various feature combinations. In FROC curves, horizontally the number of false positive detections per image is shown, vertically the sensitivity that is achieved at this specificity level. A tumor was considered detected if the peak of a detection was inside the annotation made by the radiologist. If multiple peaks were found in one annotation, they were considered as one single hit.

Peaks outside the annotated areas were counted as false positive signals. The specificity was computed using normal films only, because the presence of the tumor will have an effect of the normalized likelihood feature of normal regions in the same mammogram. Another reason to compute the specificity using only normal mammograms is that multi-focal tumors may induce false positive signals when not all tumor areas are carefully annotated or if they are missed.

The FROC-curves described above are film-based, which means that if a tumor is visible in both the oblique film and the cranio-caudal film, it is present in the set twice, and both views are treated independently. Because for many cases both oblique and cranio-caudal films were present, it was possible to compute case-based curves as well. Case-based analysis means that a tumor is considered detected if it is found in at least one of the views.

Case-based curves are always higher than film-based curves, and are more related to the way radiologists work, because in many cases a tumor is "detected" by the radiologists in only one of the available views. Except when mentioned otherwise, the curves in the results section are film-based.

Feature selection

For each of the four image characteristics that were used in this work, between 2 and 7 representing features were computed, most of them highly correlated. To select the best feature for each image characteristic, a feature selection method was required. To measure the quality of a feature, a measure similar to the A_z value in ROC-analysis was defined as the area under the logarithmically plotted FROC-curves between 0.05 and 4 false positives per image. If an FROC curve is described by $\mathrm{Froc}(x)$, the area A_f was computed by

$$A_f = \int_{0.05}^{4.0} \mathrm{Froc}(x) d\ln(x) = \int_{0.05}^{4.0} \mathrm{Froc}(x) \frac{1}{x} dx.$$

Each feature was used in combination with the 3 peak-related features to compute its A_f-value. To select the best features for the DDSM set, the neural networks were trained using the Nijmegen set, and the FROC curve was determined using this set as well. To select the best features for the Nijmegen set, the DDSM set was used for training and testing. This way, no bias is introduced by using the validation set in the feature selection procedure.

6.6 Results

The lowest lines in Figure 6.7a and 6.7b show the FROC curves that were obtained by the initial pixel-level detection step. When the normalized likelihood and the distance to the skin line were included, an improvement in performance was found. The curves that are obtained for the Nijmegen data set are higher that those for the DDSM data set, showing that the latter is a very challenging set, more difficult than the set taken from the Dutch screening.

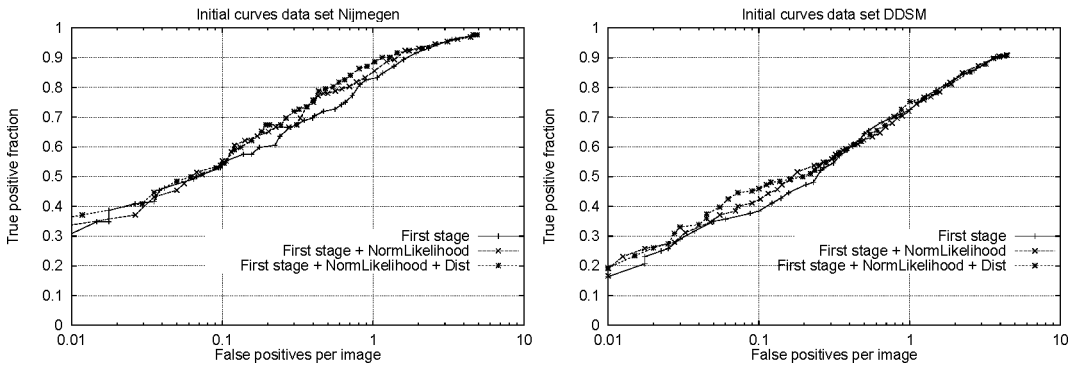

Figure 6.7: The left figure shows the results for the dataset from Nijmegen after the initial detection step and after using the three peak-related features. The right figure shows the same curves for the DDSM data set

Name	Nijmegen	DDSM
Intensity	3.532	2.907
Contrast1	3.670	2.986
Contrast2	3.702	3.129
Contrast3	3.687	3.127
Contrast4	3.675	2.987
Contrast5	3.532	2.941
Contrast6	3.577	2.946
Isodense1	3.663	3.121
Isodense2	3.604	3.058
Location1	3.541	2.938
Location2	3.560	2.940
Linear texture1	3.548	2.895
Linear texture2	3.581	2.912
Linear texture3	3.544	2.905
Linear texture4	3.587	2.923

Table 6.3: The area under the FROC curve between 0.05 and 4 was computed for each feature in combination with the 3 peak-related features. Training and testing was done on the same set.

For each image characteristic, the feature with the largest A_f-value was selected, as described in the previous section. In Table 6.3 the A_f-value for each feature in combination with the 3 initial features is presented. For both sets the same features were found, although the differences between some of the features were small. For contrast, the best feature was the normalized contrast measure (Contrast2), the best isodense feature was Isodense1, the best surrounding tissue feature was the one with the larger area and the best linear texture feature was the normalized version with the large scale. The contrast and isodense features were the best features for classification. Figures 6.8 shows the final curves when all 4 contour-related features and 3 peak-related features were included, both film-based as well as case-based results.

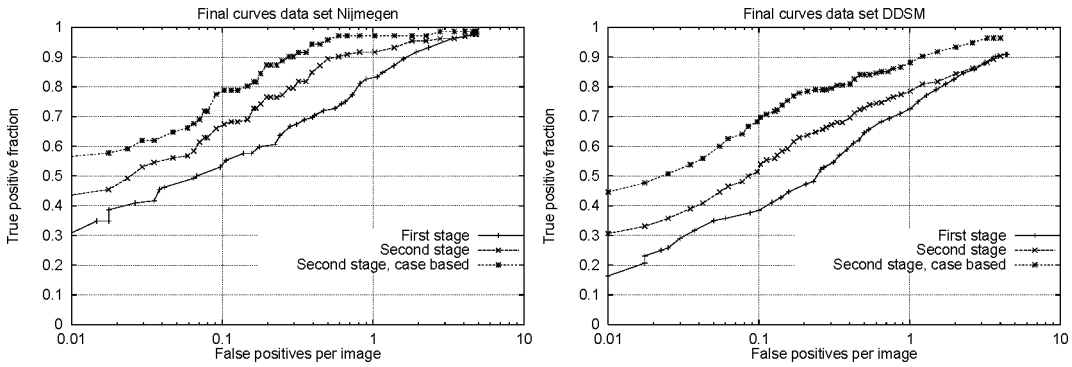

Figure 6.8: The left figure shows the results for the dataset from Nijmegen after the initial detection step and after the classification step, the latter both film-based and case-based. The right figure shows the same curves for the DDSM data set

6.7 Conclusions

A considerable improvement in performance was achieved on both data sets when a region-level classification step was applied. On the Nijmegen data set a sensitivity level close to 70% was achieved at a specificity of 1 false positive signal in 10 images. The DDSM data set is a much harder set, 55% of the masses is found at a specificity level of 1 false positive per 10 images. When case-based results are considered, the sensitivity levels become respectively 80% and 70% at 0.1 FP/image, much higher than after the initial detection step. The peak-related features improve the FROC curve considerably, and this curve was further improved by adding features representing 4 image characteristics that radiologist use. Contrast and isodense are the best features to remove false positive signals, but the linear texture feature and the location feature also contribute the the final improvement.

Bibliography

[1] SR Aylward, BM Hemminger, ED Pisano, and RE Johnston. Mixture modeling for digital mammogram display and analysis. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 305–312. Kluwer, Dordrecht, 1998.

[2] R E Bird, T W Wallace, and B C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184:613–617, 1992.

[3] K Bowyer, D Kopans, W P Kegelmeyer, R Moore, M Sallam, K Chang, and K Woods. The digital database for screening mammography. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 431–434. Elsevier, Amsterdam, 1996.

[4] H C Burrel, D M Sibbering, A R M Wilson, S E Pinder, A J Evans, L J Yeoman, C W Elston, I O Ellis, R W Blamey, and J F R Robertson. Screening interval breast cancers: mammographic features and prognostic factors. *Radiology*, 199:811–817, 1996.

[5] L Haglund, H Kuntsson, and G H Granlund. On phase representation of information. In *The 6th Scandinavian conference on image analysis*, pages 1082–1089, Oulu, Finland, June 1989.

[6] M Heath, K Bowyer, D Kopans, WP Kegelmeyer, R Moore, K Chang, and S Munishkumaran. Current status of the digital database for screening mammography. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 457–460. Kluwer, Dordrecht, 1998.

[7] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[8] N Karssemeijer and GM te Brake. Combining single view features and asymmetry for detection of mass lesions. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 95–102. Kluwer, Dordrecht, 1998.

[9] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[10] S Lobregt and M A Viergever. A discrete dynamic contour model. *IEEE transactions on medical imaging*, 14(1):12–24, march 1995.

[11] T McInerney and D Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2):91–108, 1996.

[12] W E Polakowski, D A Cournoyer, S K Rogers, M P DeSimio, D W Ruck, J W Hoffmeister, and R A Raines. Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency. *IEEE transcations on medical imaging*, 16(6):811–819, December 1997.

[13] B Sahiner, H P Chan, N Petrick, D Wei, M A Helvie, D D Adler, and M M Goodsitt. Classification of mass and normal breast tissue : a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imag*, 15:598–610, 10 1996.

[14] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[15] G M te Brake and N Karssemeijer. Automated detection of breast carcinomas not detected in a screening program. *Radiology*, 207:465–471, 1998.

[16] G M te Brake, M J Stoutjesdijk, and N Karssemeijer. A discrete dynamic contour model for mass segmentation in digital mammograms. In K M Hanson, editor, *Medical imaging 1999: Image processing*, volume 3661, pages 911–919. SPIE, 1999.

[17] B Vitak. Invasive interval cancers in the Östergötland mammographic screening programme: Radiological analysis. *European Radiology*, 8:639–646, 1998.

[18] D Wei, H P Chan, M A Helvie, B Sahiner, N Petrick, D D Adler, and M M Goodsitt. Classification of mass and normal breast tissue on digital mammograms : multiresolution texture analysis. *Med Phys*, 22:1501–1513, 9 1995.

[19] K Woods and K Bowyer. A general view of detection algorithms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 385–390. Elsevier, Amsterdam, 1996.

[20] F F Yin, M L Giger, C J Vyborny, K Doi, and R A Schmidt. Comparison of bilateral-substraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol*, 6:473–481, 1993.

Chapter 7

Comparison of segmentation methods for densities

7.1 Introduction

In Chapter 5, a number of variations of region growing and the dynamic contour model were described. Based on an overlap criterion, the dynamic contour model was found to give the best mass segmentations. In Chapter 6, a slightly modified version of this approach was used successfully to segment suspicious areas in mammograms and remove false positive signals.

The overlap criterion was useful to develop and test the segmentation methods, but the segmentations are only useful in practice when they can be used to discriminate normal tissue from real masses. It was expected that the methods that did well using the overlap criterion would also give good features for classification. In this chapter, all approaches were used to segment and classify the detected regions. The same data sets were used as in the previous chapter: the Nijmegen screening set and the DDSM data set [1].

7.2 Segmentation methods

The four versions of the region growing that were used in this work are described in Chapter 5. Two different stopping criteria were used, a simple thresholding method and a probabilistic method developed by Kupinski and Giger [2]. These two stopping criteria were used with and without the multiplication with the Gaussian to preprocess the regions. As can be seen in the left figure in Figure 5.7, preprocessing of the region with a multiplication of a Gaussian improved the performance of the region growing method.

The probabilistic method gave better results than the simple thresholding approach. In Chapter 5, the discrete dynamic contour model performed better than the region growing method. Fixed initial circle sizes were used in that chapter, the best results were obtained for circles with a radius of 8mm. In Chapter 6, the radius of the initial circle was made adaptive. The initial mass detection step located suspicious regions and generated a size estimation of the density. This estimate was used to start the method with an appropriately sized initial circle. In this chapter, the adaptive version was compared to the fixed sized version of Chapter 5.

7.3 Features

The three peak-related features described in the previous chapter were also used in this work. The four contour-related features were selected using the same method as in the previous chapter. The contours were generated using the preprocessed probabilistic region growing method.

Feature	Nijmegen	DDSM
Intensity	3.496	2.918
Contrast1	3.591	2.963
Contrast2	3.636	3.103
Contrast3	3.613	3.089
Contrast4	3.540	3.027
Contrast5	3.471	2.896
Contrast6	3.546	3.018
Isodense1	3.641	3.062
Isodense2	3.609	3.078
Location1	3.510	2.927
Location2	3.505	2.934
LinearTexture1	3.504	2.899
LinearTexture2	3.523	2.912
LinearTexture3	3.513	2.896
LinearTexture4	3.529	2.912

Table 7.1: The area under the FROC curve between 0.05 and 4 for each feature in combination with the 3 peak-related features, using the preprocessed probabilistic region growing approach. Training and testing was done on the same set.

Figure 7.1 presents the sizes of the areas under the FROC curve for the features in combination with the global mammogram features. Very similar results were obtained for this method as for the discrete dynamic contour method in the previous chapter. The only difference is that the Location2 feature was better than the Location1 feature, but these differences were so small that they were probably caused by random fluctuations. Because of the large consensus between this version of the region growing method and the dynamic contour model, feature selection using the other versions of the region growing method or the dynamic contour model was not done. Therefore, the same contour-related features were used for all segmentation methods as in the previous chapter: Contrast2, Isodense1, Location2 and LinearTexture4.

7.4 Results

The DDSM data set and the Nijmegen screening data set were used to test the mass segmentation methods. The left figure of Figure 7.1 shows that preprocessing yielded segmentations that were better to compute features on the DDSM data set. The probabilistic stopping criteria gave better results than the thresholding method. The right figure of Figure 7.1 shows

that the adaptive way to generate an initial circle worked better than a fixed initial circle size. Similar results were found for the Nijmegen data set.

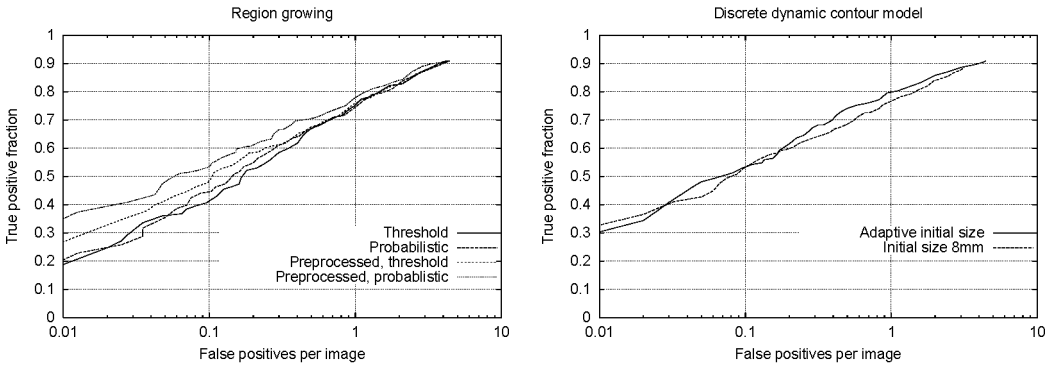

Figure 7.1: Results of the methods on the DDSM data set. Left figure: results for the four region growing methods. Right figure: results for the discrete dynamic contour model.

Figure 7.2 shows the curves of the adaptive dynamic contour model and the preprocessed probabilistic region growing method for the Nijmegen data set and the DDSM data set. The results for both methods were almost equal.

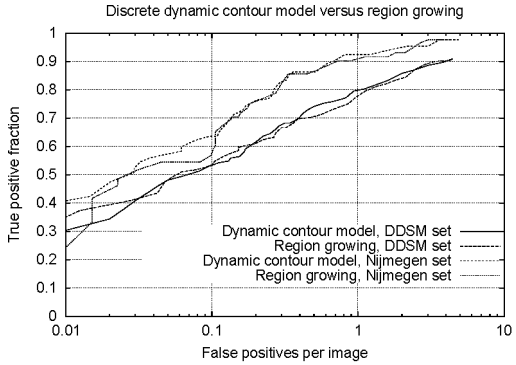

Figure 7.2: The performance of the preprocessed probabilistic region growing method versus the performance of the adaptive discrete dynamic contour model on the Nijmegen and the DDSM data sets.

7.5 Conclusions

The use of an adaptive initial circle instead of a fixed circle size improves the performance of the discrete dynamic contour model slightly. The results of the region growing method using Krupinskis' probabilistic stopping criterion and preprocessing method were close to the results of the dynamic contour model.

Bibliography

[1] M Heath, K Bowyer, D Kopans, WP Kegelmeyer, R Moore, K Chang, and S Munishkumaran. Current status of the digital database for screening mammography. In

N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 457–460. Kluwer, Dordrecht, 1998.

[2] M A Kupinski and M L Giger. Automated seeded lesions segmentation on digital mammograms. *IEEE transactions on medical imaging*, 17(4):510–517, 1998.

Chapter 8

Automated detection of breast carcinomas not detected in a screening program[1]

8.1 Introduction

In many western countries, breast cancer is the most frequently occurring form of cancer in women. For example, in The Netherlands, more than 9,600 new cases of breast cancer were detected and more than 3,500 deaths were caused by the disease in 1993 [12] . To increase the number of tumors that are detected at an early stage, breast cancer screening with mammography is becoming common in many countries. In The Netherlands, all women aged 50-70 years are invited once every 2 years to participate in a screening program. It is well known that screening mammography is difficult for radiologists and that screening errors are hard to avoid. In a previous retrospective study in Nijmegen, The Netherlands, it was found that approximately 9% of a series of cancers detected at a screening examination were visible on screening mammograms obtained 2 years earlier [15]. These were classified as screening errors. Moreover, in another 48% of the cases, a minimal sign was already visible on a prior mammogram. Of all interval cancers (cancers that appear between two screening rounds), approximately 18% were visible on prior mammograms at retrospective review; in 28%, a minimal sign was visible. Similar or even higher numbers of missed cancers have been reported in other screening programs [2, 3, 1].

These numbers indicate the importance of the development of additional tools to aid radiologists in their work, for example, a system for computer-aided diagnosis that prompts, or detects, suspicious mammographic regions. Many screening errors are perception errors. A prompt could draw the attention of the radiologist to a tumor he or she might otherwise have overlooked or to an abnormal area on a mammogram that needs careful interpretation. For this purpose, many research groups are actively designing pattern-recognition techniques to detect mammographic abnormalities. In this study, we evaluate the performance of such a system that we have developed for the automatic detection of stellate lesions [5, 6] and that has recently been extended to detect masses without spicules [14]. This system has been

[1]Published as: G.M.te Brake, N. Karssemeijer, J.H.C.L Hendriks *Automated Detection of Breast Carcinomas Not Detected in a Screening Program*, Radiology, vol. 207, pp 465-471,1998.

tested on a number of standard public databases such as the Mammographic Image Analysis Society mammographic image set [13]. In the stellate malignant abnormalities in this database, a sensitivity of more than 90% at one false-positive finding per image and a sensitivity of more than 60% at 0.1 false-positive finding per image were achieved. Here, we aim at evaluating the technique on a much more challenging database of lesions that were missed at double reading in a breast cancer screening program. This database contained consecutive series of prior screening mammograms that were classified as screening errors or minimal signs. Cases that showed only microcalcifications were excluded. All other cases, including spiculated masses and circumscribed and asymmetric densities, were included.

To our knowledge, in most publications on computerized detection of mammographic masses, a distinction is made between spiculated masses and circumscribed masses. Kegelmeyer et al [7] reported a sensitivity as high as 97% at a rate of 0.28 false-positive finding per image on a set of spiculated masses. However, these results could not be reproduced by Woods and Bowyer [16], who reported a sensitivity of only 61% at 1.4 false-positive findings per image with a different implementation of the same technique. Petrick et al [11] achieved 65% sensitivity at a rate of one false-positive finding per image on a set of circumscribed masses. In another publication on circumscribed masses, Groshong and Kegelmeyer [4] reported a sensitivity of 70% at approximately 0.6 false-positive finding per image. Nishikawa et al [9] reported a strong correlation between performance of their algorithm and tumor size. Of all tumors smaller than 15 mm, only 30% were detected at one false-positive finding per image in the study of Nishikawa et al. Of all tumors larger than 20 mm, 85% were found at one false-positive finding per image. Miller and Ramsey [8] reported that the performance of their system did not depend on the size of the tumor in a test set of screening-detected tumors. For all sizes, at a specificity level of 25% (of all women without an abnormality, 25% had a prompt in one of the mammograms (a false-positive finding on at least one of the two views)), the sensitivity was 60% (60% of the tumors were detected).

Drawing qualitative conclusions from these results is not easy, as the performance measures do not address the relative complexity of the cases that were used in the data sets. In this study, we tried to avoid this problem. All abnormal cases in our data set proved to be subtle because they were not detected by two radiologists at screening, where we assume that radiologists do not overlook obvious cancers. If a computer-aided diagnosis system is able to detect such subtle cancers with only a limited number of false-positive findings, we believe that such a system would increase the sensitivity of breast cancer screening, especially in situations where double reading is not practiced. We realize that in that respect it is important to test the sensitivity and specificity of a radiologist who is assisted by a computer-aided diagnosis system and compare the sensitivity and specificity with the performance of single and double reading. For this purpose, large, costly trials with several radiologists and a large number of normal cases are required to represent the screening situation. Before starting such trials, it is important to measure the stand-alone performance of a computer-aided diagnosis system in subtle cases. Thus, we performed this study to investigate the possibility of automated detection of early signs of cancer that were not detected in a breast cancer screening program.

8.2 Materials and methods

A data set was assembled that contained subtle tumors missed at screening. To understand the composition of the set, it is important to know that in the Dutch screening program both oblique and craniocaudal mammograms are obtained at the first visit of a woman. On succeeding visits, only oblique mammograms are obtained, unless something suspicious is seen by the radiographers, who are trained to obtain additional craniocaudal views in that case. Therefore, some cases in our set are single-view cases, while others are double-view cases.

The material used in this study consisted of series of prior screening mammograms of cancers that were detected at a later stage. All cases were reviewed by a radiologist with more than 15 years experience in breast cancer screening. Early signs of malignant lesions found in these mammograms were classified as either screening errors or minimal signs. In this classification, screening errors are cases with an abnormality that should have been detected, because clear signs of malignancy are visible in retrospect. If such signs had been detected by the screening radiologists, they would very likely have referred the case for further examination. Therefore, most of these errors are probably due to inadequate perception. Minimal signs are vague, small abnormalities that are found on prior mammograms in the region where a tumor is found at a later stage. These signs may well have been seen by the screening radiologists and considered not to be suspicious enough to recall. Therefore, the cause of not recalling cases with minimal signs is expected to be a combination of perception and interpretation error. Cases in which no sign of malignancy was found were considered to be radiographically occult and were not used in this study. It was assumed that our computer-aided diagnosis software would not generate substantial results in these cases. Also, annotation of tumor locations in these cases would be hard and very unreliable.

Assembling a large data set with missed tumors is not easy, especially if one requires that cases be recent enough to have image quality that is representative of the current state of the art in mammography. For this reason, we used a collection of mammograms of different origins. Two sources were used: prior mammograms of interval cancers and prior mammograms of screening-detected cancers. All cases were selected from screening in the cities of Nijmegen and Arnhem, The Netherlands. These cases were readily available because they are archived at our institute.

All cases of interval cancers that were used in the study were diagnosed between 1989 and 1995 in Nijmegen; there were 87 total cases. If the malignant lesion was a suspicious cluster of microcalcifications, the case was left out (four cases), as were the cases that were considered to be radiographically occult after close inspection by the radiologist (56 cases). This resulted in a total set of 27 cases with stellate lesions, circumscribed masses, and architectural distortions (20 cases with only oblique views, seven cases with oblique and craniocaudal views). Each case was classified as a minimal sign (20 cases) or as a screening error (seven cases).

The set of prior mammograms of tumors that were detected with screening is a combination of a series of cases from Arnhem and a series of cases from Nijmegen. The first part consists of 19 cases in which a minimal sign was found in retrospect on a screening mammogram obtained 2 years before detection. These cases were imaged between 1992 and 1994 in Arnhem and are all shown on single views. The second part consists of cases taken from a series of prior screening mammograms of screening-detected tumors that were larger

than 15 mm at detection. These cases were imaged in Nijmegen between 1990 and 1993. After exclusion of radiographically occult cases and cases with only microcalcifications, 19 cases were left, of which 16 had only oblique views and three had oblique and craniocaudal views. Every case in these two series of screening-detected cancers was classified as a minimal sign (31 cases) or as a screening error (seven cases).

For the screening-detected cases, the mammograms at the stage of detection also were digitized to compare the performance of the software on these mammograms with the performance on the mammograms obtained 2 years before. A total of 47 cases (11 masses, 33 spiculated lesions, and three architectural distortions) were present in this set. This is larger than the 38 cases classified as minimal signs or screening errors, because cases classified as occult at the prior screening also were included. Mammograms at the stage of detection in the series of interval cancers were not used, because for most of these cases only copies were available in our institute.

For estimation of the specificity of detection algorithms, that is, the number of false-positive findings per image, there should be enough normal mammographic tissue in the mammograms in the study set. Therefore, we included all contralateral mammograms available in the cases that we selected: a total of 142 mammograms without abnormalities. It was verified that no abnormality was found on these mammograms in succeeding screening rounds.

To construct the detection schemes, a training procedure is required with mammograms of known malignant cases. The training set we used contained 36 mammograms obtained from two institutions. Ten malignant architectural distortions, eight malignant spiculated lesions, and four malignant circumscribed masses were included from the Mammographic Image Analysis Society database [13]. In addition, we included 14 mammograms containing malignant stellate lesions that were digitized in Nijmegen. The mammograms that were used to train the detection schemes were not used in the other sets for evaluation of the schemes.

An overview of the size and contents of the image sets that were used in this study is given in Table 8.1. All shown numbers relate to the number of images; for the number of cases, refer to the text.

	Films	Masses	Spic. lesions	Arch. distortions	Screening errors	Minimal signs
Interval	34	20	6	8	9	25
Screening	41	15	17	9	9	32
Training	36	4	22	10	n.a.	n.a.
Detection	73	16	54	3	n.a.	n.a.

Table 8.1: Composition of the sets. All shown numbers are number of films.

The screening-detected cases from Arnhem and the 14 training images that were digitized in Nijmegen were digitized with a resolution of 100 m and 12 bits per pixel by using an Eikonix 1412 CCD camera (Eastman Kodak, Rochester, NY). The images in the interval set and both the prior images and the detection images of the screening-detected cancers from Nijmegen were digitized at a resolution of 50 m by using a model 85 digitizer (Lumisys, Sunnyvale, Ca.). All images were averaged down to 200 m to reduce the computation load. To detect masses and stellate lesions, images with a resolution of 50 m per pixel are probably

not required. The images from the Mammographic Image Analysis Society database were digitized with a Scandig-3 scanning microdensitometer (Joyce-Loebl Automation, Sunderland, England) at 50 m and 8 bits. These images were averaged down to 200 m per pixel and 12 bits. All tumors were annotated by an experienced radiologist.

Three different detection schemes were applied to the sets of mammograms. The first scheme detects radiating patterns of linear spicules. If a clear mass is present without spicules, it will not be found. However, subtle radiating architectural distortions without a mass may be found at high specificity levels. The second scheme detects bright regions. It is sensitive for clear, circumscribed masses but will not detect spiculated distortions without a central mass. The third scheme is a combination of the first two. It incorporates both a component that detects masses and a component that detects spicules. Subtle, lightly spiculated masses may be found only at acceptable specificity levels by a system that combines these two properties.

In an additional experiment, the schemes were extended with the analysis of local oriented edges (ALOE) method [7], which has been reported to detect stellate abnormalities with a very high sensitivity. Like our own approach, this method is based on the detection of local disruptions in edge or line orientation patterns in a region of the breast. One crucial difference is that the ALOE is based on absolute orientations, while our approach is based on orientations relative to a central point. This allows us to define measures that are directly related to the amount of spiculation around a region.

All schemes that were used assign a measure of suspiciousness to each pixel in the image on the basis of the presence of spicules, masses, or both in the surrounding region. For each pixel, five "image features" are computed, numeric values that describe local image properties. Two features will have high values at the center of a radiating pattern of spicules, two other features will have high values for pixels inside a bright region. The ALOE feature will have a low value if the global tissue structure is disrupted in the region surrounding the pixel. The detection schemes contain a neural network that assigns a measure of suspiciousness to each pixel on the basis of the numeric values of the features it uses. This set of numeric features is called a feature vector. The neural network has to be trained to "learn" the differences in the feature values between pixels in normal and abnormal tissue. By showing it a large number of feature vectors belonging to pixels in normal tissue and another set of feature vectors representing pixels inside annotated malignant regions, the neural network learns to map a feature vector to a measure of suspiciousness. On the basis of the feature vectors that have been presented to the neural network during the learning phase, it can classify new feature vectors from mammograms that were not in the training set. A more elaborate technical explanation of the system is given in the Appendix and in reference [6].

When the neural network has assigned a measure of suspiciousness to each pixel, a threshold can be applied to find regions that are suspicious to some degree (Fig 8.1). In an application, these could be signaled to the radiologist. By lowering this threshold, a more sensitive system is obtained. By raising this threshold, fewer regions will be signaled, giving improved specificity at the cost of lower sensitivity. By varying the threshold, the performance of the detection systems can be measured as free response operating characteristic curves. In these plots, the sensitivity in the test set is given as a function of the average number of false-positive findings per image or per case. We use a logarithmic scale on the horizontal axis to emphasize the performance at higher specificity levels.

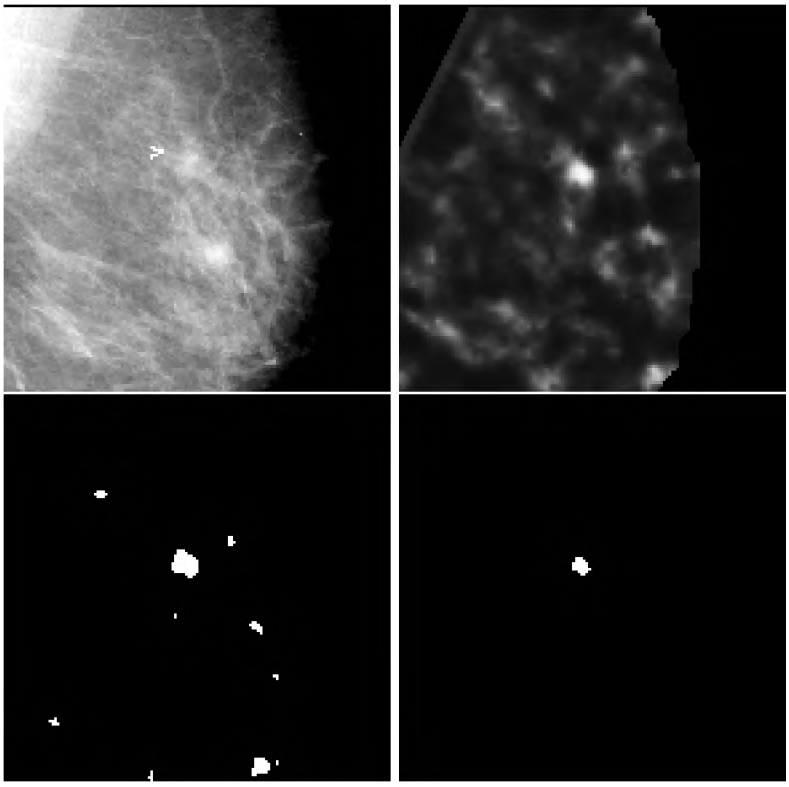

Figure 8.1: Top left: Part of a mammogram containing a stellate lesion (arrowhead) classified by the radiologist as a minimal sign. Top right: the output of the neural network for this region (white indicates suspicious of malignancy). Bottom: Results of two different thresholds levels: left at approximately 3 false positives per image, right at a specificity level of 0.15 false positives per image.

To construct free response operating characteristic curves, clear definitions of true- and false-positive findings are required. A variety of criteria can be used to determine whether a tumor has been detected Here, we considered a tumor detected if a region was classified as suspicious and if its most suspicious point was inside the region annotated by the radiologist. Otherwise, a false-positive finding was counted. This is a strict criterion that proved to be relatively independent of the size of the annotations.

If a suspect mammographic region is detected by a radiologist on only one view, the woman will likely be called back for clinical examination, regardless of a second view, which may reveal no sign. Therefore, if a tumor is detected on at least one view by our detection system, one might argue that the tumor should be considered detected. Results obtained this way will be referred to as case based. In mammogram-based free response operating characteristic curves, hits and misses are counted for each view separately when two views from the same breast are present. Generally, case-based curves present a more optimistic view on the results when many two-view cases are present.

8.3 Results

In Figure 8.2, mammogram-based results are shown for the three schemes applied to the set with the prior screening images of interval carcinomas. Figure 8.3 shows mammogram-based results for the set with the prior screening images of screening-detected carcinomas. In each experiment, all normal cases were included. It appears that the scheme that includes both mass and spicule detection has a similar performance on both sets. This is not the case for the other two schemes. The mass scheme has a better performance with interval cancers, while the spicules scheme has a better performance with the screening-detected cancers. The reason for this is the composition of the sets (Table 1).

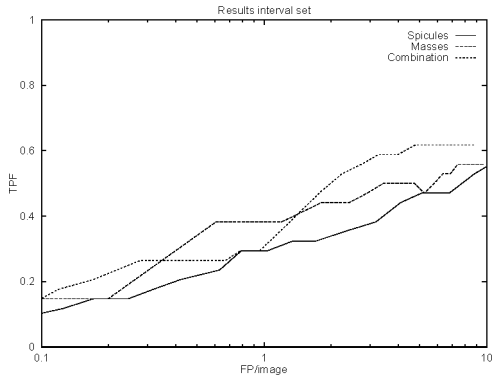

Figure 8.2: Mammogram-based FROC curves representing the performance of the automated detection schemes on previous screening mammograms of interval carcinomas.

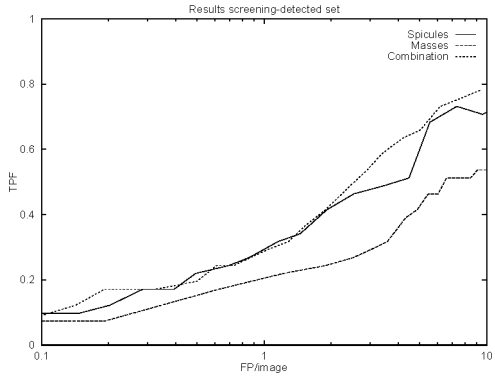

Figure 8.3: Mammogram-based FROC curves representing the performance of the automated detection schemes on previous screening mammograms of screening-detected carcinomas.

Both sets were combined to make a set of 65 cases with signs of cancer that were not detected at screening. In Figure 8.4, results are shown for the three schemes with this combined set. This curve is case based. It appears that the combination scheme slightly outperforms the other two schemes. For comparison, we applied the schemes to the images in which the tumor was discovered. Because many original images of the interval carcinomas were not available, only the screening-detected cancers (47 cases) were used in this experiment [10]. The results in this set are shown in Figure 8.5.

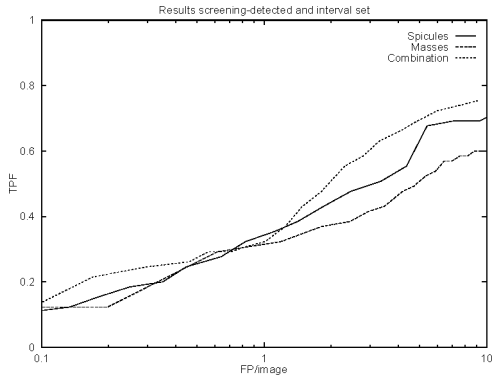

Figure 8.4: Case-based FROC-curves of the three detection schemes on previous screening mammograms of interval- and screening detected carcinomas.

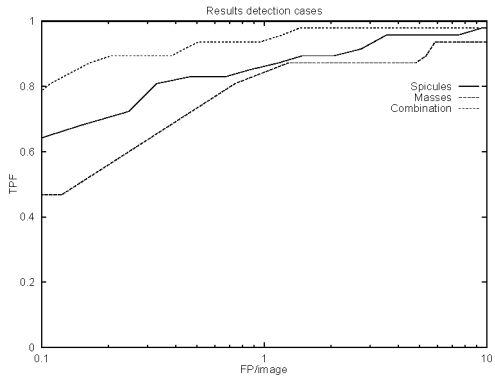

Figure 8.5: Case-based FROC-curves of the three detection schemes of the screening detected cases at the detection stage.

The abnormalities of the screening-detected set and the interval set were partitioned into two sets–minimal signs and screening errors–by radiologists with breast cancer screening experience. Fourteen cases were classified as screening errors; 51 cases were classified as minimal signs. In Figure 8.6, we show the performance of the scheme that detects both spicules and masses on the two sets. As expected, tumors that were classified as screening errors are detected with higher specificity levels than tumors that were classified as minimal signs.

All tests were repeated with ALOE as an additional feature. The curves were similar to the curves obtained without this feature. No improvement was found in any of these experiments.

8.4 Discussion

The results show that a substantial number of tumors that were missed in a screening program for breast cancer, despite double reading, are found at specificity levels of less than one false-positive finding per image by our automated detection system. The system that detects both masses and spicules outperforms the basic schemes of the software for most cases.

At one false-positive finding per image, the system that detects both masses and spicules

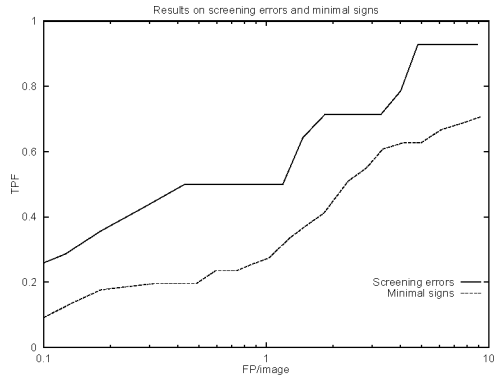

Figure 8.6: Performance of the system on cases classified as screening errors and cases classified as minimal signs.

correctly marked tumors in at least one view on prior screening mammograms in 22 (34%) of 65 cases. At three false-positive findings per image, a sensitivity of 60% (39 of 65 cases) was achieved. Of the 14 cases classified as screening errors, seven (50%) were found at a specificity of less than 0.5 false-positive finding per image, and a sensitivity of 71% (10 of 14 cases) was reached with less than two false-positive findings per image. Although it is not clear what specificity is required for a computer-aided diagnosis system to be effective, these numbers indicate that a substantial number of missed cancers could have been detected if a system such as ours would have signaled early malignant signs to the screening radiologists. Large-scale experiments need to be set up to test this hypothesis and to study the effect of false-positive findings on the specificity of radiologists.

Surprisingly, addition of the ALOE feature that forms the basis of a very successful method reported in the literature did not increase the performance of our schemes on any of the test sets. This means that the neural network classifier we used was not able to extract any useful additional information from this feature. Although this is puzzling in light of the results reported by Kegelmeyer et al [7], it is in accordance with results published by Woods and Bowyer [16], who reported poor performance of their implementation of Kegelmeyer's algorithm on a set of 320 mammograms containing 62 stellate lesions.

Good results were obtained on the screening mammograms obtained at the stage of detection. A case-based sensitivity of 80% (38 of 47 cases) was achieved at 0.1 false-positive finding per image, and a case-based sensitivity of approximately 96% (45 of 47 cases) was achieved at 1.2 false-positive findings per image. This performance far exceeds the results reported by most other groups. However, it should be noted that many of these cancers were masses larger than 1.5 cm.

Tumors that were classified as screening errors were on average found at better specificity levels than tumors that were classified as minimal signs. However, some minimal signs were signaled with a very high measure of suspiciousness. This indicates that tumors that are hard for radiologists to detect are not necessarily the hardest for an automated detection system.

The difference between case-based and mammogram-based curves is not large. This is due to the fact that only a few two-view cases are present in the sets. This is normal in the Dutch screening, because two views are obtained only at the first visit for screening or when the radiographer expects that the one oblique view is not sufficient for the radiologists be-

cause of breast composition or a suspicious area. The case-based curves give a slightly more positive statistic than do the mammogram-based curves. There is especially a difference in spiculated abnormalities, because they often are visible only on one view, whereas masses often are visible on both views.

Many of the results published in digital mammography are achieved with a single data set in which all mammograms have been digitized identically. For our experiments, mammograms were digitized at different institutions, at different resolutions, and by using different equipment. Despite this variation in the acquisition of the digital mammograms, the system yielded good results. This shows that the detection system is robust and increases confidence in the general validity of our results.

Acknowledgment: We thank Daniel J. Dronkers, MD, PhD, for providing the data on the mammograms collected in Arnhem, The Netherlands.

Appendix

The images that are used in the algorithm have a resolution of 200 m per pixel. This gives images of approximately 700 1,000 pixels, of which approximately 50% are inside the breast area, depending on the size of the breast. The approach that is used in our method is pixel based. Every pixel is assigned a measure of suspiciousness on the basis of some local signs, such as location inside a mass or being surrounded by a radiating pattern of lines.

Quantitative measures indicating the presence of these signs are defined and are referred to as features. These features are mapped to a measure of suspiciousness by a classification system. An elaborate description of these features can be found in an article by Karssemeijer and te Brake [6]. A brief description of the way the features are defined is given here. The approach is based on analysis of local orientation patterns. On a mammogram, many line structures are visible. Some of these may represent spicules of a tumor, but the vast majority will reflect normal parenchymal patterns and blood vessels. In our method, at each pixel a line orientation estimation is carried out by using a small local neighborhood, which is accurate if some linear structure is present and gives a random orientation elsewhere. The resultant map of orientations forms the basis for the detection of stellate structures. If a pixel is located in the center of a radiating pattern of spicules, the number of nearby points n with an orientation toward this central pixel is high. If no structure is present, we can estimate the expected number of pixels that are randomly oriented toward this central pixel. We can also estimate the variance of this number. Our first feature, f_1, is defined as $f_1 = (n - n_{exp})/(n_{var})^{\frac{1}{2}}$, where n_{exp} is the expected value of n and n_{var} is the variance of n. All pixels in a circular neighborhood of a few centimeters in diameter contribute. The circular neighborhood is divided into a number of bins. If a high number of pixels oriented toward the center is present in many bins, this is more suspicious than when all these pixels are concentrated in only a few bins. Our second feature expresses this. The number of expected hits, that is, pixels with the right orientation, is computed for each bin. Let l be the number of bins with more hits than expected and k be the total number of bins; our second feature, f_2, is defined as $f_2 = (l - k/2)/(k/4)^{\frac{1}{2}}$.

A neural network has been trained on these two features by using a set of pixels taken from the training set. A feed-forward network was used with two inputs, five hidden nodes,

and two output nodes (the first giving a high value if the pixel is normal, the other if it is malignant). The difference between the two output nodes is used as the measure of suspiciousness. This neural network is the scheme in this article that detects only spicules.

A third and a fourth feature to represent the presence of a mass were defined in a similar way as the first two features. For each pixel, the intensity gradient is computed by using a small local neighborhood, as well as the orientation of this gradient. If a pixel is lying inside a mass, there will be many points in its neighborhood with a gradient orientation pointing away from this central pixel. Two features representing the number of pixels with this property and their distribution over the bins are computed in the same way as for the first two features. A neural network is trained by using a set of feature patterns with these two features to create a system for mass detection. Again, it has two input nodes, five hidden nodes, and two output nodes. This neural network is the scheme that detects circumscribed masses.

The combined automated detection scheme for masses and stellate lesions is a neural network with four input nodes, five hidden nodes, and two output nodes. All four features were used to train the network and to classify pixels from the test mammograms. For the experiments where the ALOE feature was included, an additional input node was created, and all other factors were kept the same.

Bibliography

[1] R E Bird, T W Wallace, and B C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184:613–617, 1992.

[2] H C Burrel, D M Sibbering, A R M Wilson, S E Pinder, A J Evans, L J Yeoman, C W Elston, I O Ellis, R W Blamey, and J F R Robertson. Screening interval breast cancers: mammographic features and prognostic factors. *Radiology*, 199:811–817, 1996.

[3] S Ciatto, M Roselli del Turco, and M Zappa. The detectability of breast cancer by screening mammography. *Br. J. Cancer*, 71:337–346, 1995.

[4] B R Groshong and W P Kegelmeyer. Evaluation of a hough transform method for circumscribed lesion detection. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 361–366. Elsevier, Amsterdam, 1996.

[5] N Karssemeijer. Recognition of stellate lesions in digital mammograms. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 211–220. Elsevier, Amsterdam, 1994.

[6] N Karssemeijer and G M te Brake. Detection of stellate distortions in mammograms. *IEEE Trans Med Imag*, 15:611–619, 10 1996.

[7] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[8] L Miller and N Ramsey. The detection of malignant masses by non-linear multiscale analysis. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 335–340. Elsevier, Amsterdam, 1996.

[9] R M Nishikawa, M L Giger, K Doi, C J Vyborny, and R A Schmidt. Computer-aided detection and diagnosis of masses and clustered microcalcifications from digital mammograms. In K W Bowyer and S M Astley, editors, *State of the art in digital mam-*

mographic image analysis, volume 9 of *Series in machine perception and artificial intelligence*, pages 82–102. World Scientific, 1994.

[10] D Chakraborty P and L Winter L H. Free response methodology: Alternate analysis and a new observer-performance experiment. *Radiology*, 174:873–881, 1990.

[11] N Petrick, H P Chan, B Sahiner, and D Wei. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. *IEEE Trans Med Imag*, 15:59–67, 1996.

[12] The Netherlands Cancer Registry. Incidence of cancer in the netherlands. Report, 1993.

[13] J Suckling, J Parker, D R Dance, S Astley, I Hutt, C R M Boggis, I Ricketts, E Stamatakis, N Cerneaz, S L Kok, P Taylor, D Betal, and J Savage. The mammographic image analysis society digital mammogram database. In A G Gale, S M Astley, D R Dance, and A Y Cairns, editors, *Digital Mammography*, pages 375–378. Elsevier, Amsterdam, 1994.

[14] G M te Brake and N Karssemeijer. Detection of stellate breast abnormalities. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 341–346. Elsevier, Amsterdam, 1996.

[15] J A M van Dijck, L M Verbeek, Hendriks J H C L, and R Holland. The current detectability of breast cancer in a mammographic screening program. *Cancer*, 72:1933–1938, 1993.

[16] K Woods and K Bowyer. A general view of detection algorithms. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 385–390. Elsevier, Amsterdam, 1996.

Chapter 9

Experiments with a computer aided diagnosis system

9.1 Introduction

To reduce the number of errors in breast cancer screening programs, automated detection systems have been built that signal suspicious regions using prompts. If prompts are generated by a system at a high sensitivity level and all prompted areas are examined by the radiologist, the number of errors due to oversight will diminish. Early work in mammography by Chan [3] and Kegelmeyer [6] showed that prompts improved radiologists' performance for detection of microcalcifications and spiculated malignancies. In these two studies, the percentage of abnormalities was respectively 50% and 40%. In a few recent experiments with prompted radiologists, a lower target rate of approximately 5% was used [9, 12]. These experiments showed that the specificity of readers using a CAD system is equal to the specificity of readers reading in the normal way. In both studies, the recall rate was about 8%, much higher than normal in screening programs. In the Dutch program, the recall rate is approximately 1%, much lower than in other countries and in these experiments.

In 1997, a commercially available prompting system was placed in our radiology department, creating the opportunity to get practical experience with a prompting system. The CAD system that was used in the experiments was R2's ImageChecker. The ImageChecker consists of two parts: a unit where the mammograms are digitized and examined by a computer program, and an alternator with two small built in monitors where the computer findings can be shown to a radiologist examining the case. On this monitor a low resolution image of the mammograms is visible with prompts showing possible masses and microcalcifications. The ImageChecker is designed to prevent errors due to oversight. At the sensitivity level at which the system operates a rather large number of false positives are generated (approximately 1 prompt per image).

The presence of the system was used to do two studies. There are worries that the low specificity level of the CAD-device has a negative effect on the specificity of screening radiologists. To investigate this, an experiment was designed to examine the effect of prompts on the recall rate of the radiologist. The effects of the prompting system on the sensitivity of the radiologist could not be studied in detail in this experiment, because no recordings were made whether prompts were used or ignored. Screening forms similar to those used in practice were used.

A set of 600 cases containing a large number of normal cases to reflect the screening situation was constructed. Twelve experienced screening radiologist read 300 cases, 200 in the normal way and 100 assisted by the CAD-system. The types of signs that cause false positives and the types of abnormalities that cause false negatives were examined as well.

In a second experiment, the potential benefit of prompting for masses was studied. Eleven radiologist were asked to mark all the regions they examined in a mammogram, and grade these on a 6-point scale. A monitor was attached to the alternator, showing the case that was read. With a mouse, the radiologist pointed out all areas they examined for being a possible malignant mass, which enables us to separate interpretation errors and detection errors. A set of 120 cases was composed with 13 malignant masses.

In both experiments, the variation in performance of radiologist was examined, because some studies suggest that all radiologist work on the same ROC curve [8] and others that rather large variations in performance between radiologists exist [10, 1, 5].

9.2 Experiment 1: specificity of prompted radiologists

9.2.1 Materials and methods

The cases for the experiment were taken from the Dutch screening program, which at that time covered the age group 50-69. Screening is done bi-annual with a four-view mammogram at the first visit and is restricted to oblique views at follow up. However, when radiographers suspect an abnormality or judge the films hard to read because of dense tissue, additional cranio-caudal views are made. Four views are made in 10 to 20 percent of these followup cases, and this percentage appears to increase over the years.

Constraints on the availability of the readers in the experiment enforced us to limit the amount of cases read by each radiologist to 300. To increase the variation of cases we decided to construct a set of 600 cases, of which half would be assigned to each radiologist attending the experiment. The total number of films in the set was 1428 (1200 oblique and 228 cranio-caudal). These mammograms were taken from screening Round 10 (1993/1994) in the region of Nijmegen, only original films were used. In addition to the films from Round 10, mammograms from the previous screening were presented in the reading sessions, but these were not processed by the CAD system. To compose an appropriate set, six types of cases were distinguished:

Negative, type 1: Cases that were found negative in Round 10, and were also found negative in Round 11 (1995/1996).

Negative, type 2: Cases that were suspected positive in Round 10, but were found negative after work-up (usually proven by biopsy).

Negative, type 3: Cases that were found negative in Round 10, but turned out to have a cancer in or before Round 11 that in retrospect was judged not visible in Round 10 (radiographically occult).

Negative, type 4: Cases that were found negative in Round 10, but turned out to have a cancer in or before Round 11 where in retrospect a minimal sign (not actionable) was visible in Round 10.

Positive, type 1: Cases that were found positive in Round 10, and were proven malignant indeed by biopsy.

Positive, type 2: Cases that were found negative in Round 10, but where a cancer was found in or before Round 11 that in retrospect was already visible in Round 10 ("screening error").

To create a sample of cases that realistically reflects problems encountered in screening, it was chosen to represent all types of cases in the data set. It would have been desirable if the distribution of these 6 types was similar to that in real screening, but this would mean that only 2 or 3 positives would be included. As a trade-off, 33 positives were included in the set. In Table 9.1 an overview of the composition of the set is given.

Case type	% in screening	% in study	Number in set
Negative, type 1	99.23	89.0	534
Negative, type 2	0.19	2.5	15
Negative, type 3	0.10	2.2	13
Negative, type 4	0.07	0.8	5
Positive, type 1	0.37	4.0	24
Positive, type 2	0.04	1.5	9
Totals	100%	100%	600

Table 9.1: Composition of the set

In Table 9.2 an overview of the types of malignant and benign abnormalities and the number of occurrences is given. The malignant abnormalities are the two positive types, the other abnormalities are of negative type 2 (benign) and negative type 4 (not actionable). Some very obvious positive cases were left out, because these would not contribute to the statistical power of the test. It is possible that in the other negative types abnormalities were present, which were not found or considered benign in Round 10 and Round 11, but the chance that this is the case is small.

The order of the 600 cases was randomized, after which the set was divided into six subsets of 100 cases. Therefore, the number of abnormal cases was not equal in all sets. One set did not include any malignant abnormalities, the other sets included between 5 and 10 malignant abnormalities. The benign abnormalities were also distributed randomly over the 6 sets.

In Table 9.3, the sensitivity of the CAD-system is shown per type of malignancy on the set that was used in this study. In the two cases where both a mass and microcalcifications were visible the lesions were found only by the microcalcification detection software. The total number of markers indicating a possible spiculated lesion was 959 (0.65 per film), of which 19 were correct (12 cases). The total number of markers indicating microcalcifications was 756 (0.51 per film), of which 20 were correctly hitting a malignant cluster (9 cases).

Twelve radiologists participated in the experiment, all certified to read mammograms and involved in breast cancer screening in the Netherlands. Although the amount of experience varied among the radiologists, they all read at least 2000 mammograms each year for at least 3 years. The radiologists read in two different modes:

	Positive	Negative		
Type of abnormality	Both types	Type 2	Type 4	Total
Spiculated mass	9	1	1	11
Architectural distortion	4	1	2	7
Nodular mass	2	8	0	10
Mass with vague margins	9	2	1	12
Microcalcifications	7	3	1	11
Spic. mass + m.c.	1	0	0	1
Vague mass + m.c.	1	0	0	1
Totals	33	15	5	53

Table 9.2: Overview of the abnormalities

Type of abnormality	#	CAD sensitivity
Spiculated masses	9	66.67%
Architectural distortions	4	100.00%
Nodular masses	2	0.00%
Vague masses	9	22.22%
Microcalcifications	7	100.00%
Spic. mass + m.c.	1	100.00%
Vague mass + m.c.	1	100.00%
Totals	33	63.64%

Table 9.3: Sensitivity of the CAD-system by type of malignancy.

Normal reading: A radiologist reads cases on a conventional alternator, and decides whether a follow-up examination is required.

CAD-assisted reading: A radiologist reads films on the R2-alternator and examines the films. Next, he activates the markers by pressing the button and checks whether a potential abnormality flagged by the system has been overlooked. Then a final decision is made by the radiologist. What the decision was before the prompts were provided was not recorded.

For each case the radiologists filled in a form that was similar to the ones used in screening. Using a five category scale each case was reported as normal (1), mastopathy (2), abnormal benign (3), probably malignant (4), or certainly malignant (5). The term mastopathy is used for fybrocystic changes in the breast that make mammograms very dense and hard to read. Cases in category 4 or 5 are referred for work-up. It is noted that this scale does not allow ROC analysis, which is a clear disadvantage. However, it was chosen to keep as close to screening practice as possible. For all cases rated in category 3, 4, and 5 the radiologists had to report the type of finding and its location. The following types were distinguished: microcalcifications, masses, mass and microcalcifications, architectural distortions and other abnormalities.

Each participating radiologist read 3 different sets, two as a normal reader, one CAD-assisted. All sets were read the same number of times, and were equally spread over the two

reading modes.

9.2.2 Results

To determine the specificity of the various reading modes, only negative type 1 cases were used. Negative type 2 cases (proven benign abnormalities) were left out because referral of most of these cases cannot be considered as a wrong decision. Negative type 3 and 4 were not taken into account in the main analysis of sensitivity and specificity because classification of these cases is rather subjective. Results for these types were analyzed separately. The sensitivity was computed using both positive types. The results of the 12 radiologists were pooled to compare the two reading modes. For each radiologist, the results in the two reading modes are pooled to compared his performance to the other radiologists.

In the next sections, the effect of prompts on the specificity of a screening radiologist is described, the inter-observer variability is examined, and the false positive and false negatives are examined.

Effects of prompts on the specificity of a radiologist

In Table 9.4, the sensitivity and specificity of the readers for the two reading modes are shown. To investigate the correlation between the decisions made by the radiologist and the possibilities for improvement by double reading, two additional reading modes were computed. Two radiologist who read the same set were coupled. In double mode 1, the case was only submitted if both reader grades it 4 or 5, in double mode 2 if at least one found it suspicious.

Reading mode	Sensitivity	Specificity
Normal	76.52%	97.28%
CAD-assisted	66.67%	98.21%
Double 1	62.12%	99.91%
Double 2	90.91%	94.66%

Table 9.4: Sensitivity and specificity for the two reading modes, and the two computed double modes.

The results in Table 9.4 show that the specificity for normal and CAD assisted reading is similar. The numbers were computed by pooling the results of the 12 radiologists using all 6 sets. The false positive markers of the CAD system did not decrease the specificity of the observers, the specificity was even a little higher than for the normal reading mode. The number of of true negative findings is binomially distributed. We can assume there is no difference in specificity for the two reading modes, and that the probability p that a negative case was not referred is constant and equal for both reading modes. Lets estimate that this probability p is 0.975 for both reading modes. The number of readings of negative cases in the normal reading mode was n=2136, for the CAD-assisted reading n=1068. Using the definition of the variance of the binomial distribution we can compute a standard deviation of 0.35% for the normal reading mode and 0.49% for the CAD reading mode. For this computation, the readings of normal cases are assumed to be independent. This is a

reasonable assumption because the specificity values of both modes are very close to the expected values if independence was assumed. Both found specificity values are inside the 95% confidence interval of the mean specificity of 97.5%, so there is no reason to reject the hypothesis that they are equal with this probability value.

Inter-observer variability of radiologists

We cannot directly compare the individual performance of radiologists, because they did not read the same sets. Not all sets were equally hard, variations were found in the average achieved sensitivity and specificity of the sets, some radiologists were assigned more difficult sets than others. For each set, the average performance was computed. The performance of the radiologists could be compared to the average performance on his 3 sets. In Table 9.5 the performance of each radiologist is given, as well as a normalized score based on the complexity of his sets. The corrected value show the difference between the achieved score and the average score on the 3 sets that were read by the radiologist. Figure 9.1 shows the corrected performance of each of the radiologists in a scatter plot.

Radiologist	Achieved sensitivity	Achieved specificity	Corrected sensitivity	Corrected specificity
A	66.67%	98.87%	-13.33	1.74
B	60.00%	99.63%	-14.00	2.34
C	23.08%	99.25%	-43.07	2.25
D	55.00%	96.24%	-22.00	-1.65
E	91.30%	98.08%	20.87	0.70
F	73.33%	98.51%	-8.00	0.67
G	65.22%	91.09%	-0.87	-7.05
H	100.00%	98.17%	34.00	0.21
I	80.00%	98.88%	-3.00	1.29
J	100.00%	98.91%	12.00	0.87
K	86.96%	95.31%	15.66	-2.22
L	76.92%	97.75%	21.54	0.45

Table 9.5: Achieved and corrected sensitivity and specificity for each radiologist. Corrected values are computed by subtracting the average score on the sets from the achieved score.

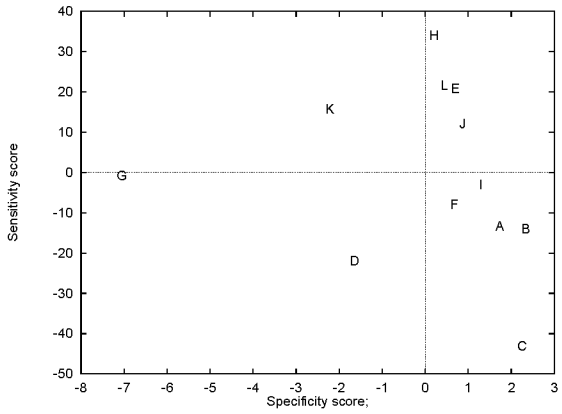

Figure 9.1: Sensitivity and specificity for each radiologist.

False negatives and false positives

For each case, the radiologist filled in a report form, which enables us to some extent to discriminate perception errors (lesions missed due to oversight) from interpretation errors (lesions that were detected but judged benign). The R2 ImageChecker, as discussed above, is intended to avoid perception errors.

Examination of not recalled malignant tumors showed that at least 50% can be accounted to interpretation errors. Combining all sessions of the twelve radiologists, in the normal mode and the CAD-assisted mode, each case was read six times. In total, 53 times a malignant case was not sent for further examination. These were reported as benign 27 times, and as normal 26 times. The first were clearly errors due to interpretation: the abnormality was examined and classified benign. The latter will partly be due to interpretation errors and partly to detection errors. In Table 9.6 an overview per type of malignancy is given. Architectural distortions and microcalcifications were not recalled more often than the other abnormalities. The distortions were not referred in 45% of the readings, the malignant microcalcification clusters were found negative in 35% of the readings (Table 9.6). It is noted that the sensitivity of the CAD system for architectural distortions and microcalcifications was 100% on the test set. Assuming that the radiologists examined all prompted areas, it is likely that these areas were examined but not found suspicious enough to be recalled. In other words, most of these errors should be classified as interpretation errors.

In Table 9.7, the reason for referral of negative cases is shown. Of all mammograms of negative type 1 (the normals that were used to compute the specificity), 77 times a case was found suspicious, most often because a mass was seen. It appeared that one particular case was recalled 5 times out of the six times it was read. A small mass was visible in this case, which was verified to be normal by follow up after two years. Furthermore, it was found that two cases were recalled three times, seven cases were recalled twice, and 57 cases once. A similar analysis was performed for the other negative types. Negative type 3 cases were recalled more often than the Negative type 1 cases (9% versus 2.5%). This difference in recall rate suggests that mammograms that were classified occult by the radiologist that helped composing the set, showed suspicious signs to other radiologists.

It appeared that the sensitivity of the two reading modes for the malignant masses is high. In the normal reading mode, approximately 80% of the malignant masses were referred for

	Readings	Not referred		Reason	
Malignant sign	#	#	%	Unknown	Interpretation
Spiculated mass	54	13	24.1	7	6
Arch. distortion	24	11	45.8	6	5
Nodular mass	12	4	33.3	1	3
Vague mass	54	9	16.7	4	5
Microcalcifications	42	15	35.7	8	7
Spic. mass + m.c.	6	1	16.7	0	1
Vague mass + m.c.	6	0	0	0	0
	198	53	26.8	26	27

Table 9.6: Positive cases not referred for further examination.

	Readings	Referrals		Reason for referral				
	#	#	%	?	m.c.	Mass	Mass +m.c.	Arch. Dist.
Negative type 1	3197	77	2.5	9	6	45	4	13
Negative type 2	90	36	40.0	1	7	26	1	1
Negative type 3	78	7	9.0	2	4		1	
Negative type 4	30	11	36.7	4	1	3	1	2
Totals	3395	131	3.9	16	18	74	7	16

Table 9.7: Negative cases referred for further examination. For each type, the number of times it is referred, and the reason why it was referred by the radiologist are given.

further examination. However, this sensitivity was reached at the cost of 74 false positive readings of masses, of which 45 occurred in negative type 1 mammograms (i.e. cases without any findings). The sensitivity for microcalcifications is lower, approximately 65%, but only 6 times a negative type 1 case was sent for further examination because suspicious microcalcifications were seen. The different performance on microcalcifications seems to reflect the less aggressive approach radiologists in the Netherlands are trained to practice, to avoid too many false positive referrals.

9.3 Experiment 2: interpretation of mass-like regions

In the previous experiment each radiologist recalled a number of cases, true as well as false positive cases. A large number of areas in these and other mammograms were examined as well but found not suspicious enough to recall the case, but this was not recorded. A second experiment was designed to give insight in the decision process of the radiologist examining masses and therefore in the potential benefit of prompting for masses. Several studies have shown that a high percentage of missed malignancies are masses, and also a large number of false positives are of this type [11, 2]. In this experiment, radiologist were asked to examine a set of cases and indicate all areas in the mammograms that they examined for the presence of signs of a malignant mass. Because all areas that were examined were pointed out by the

radiologist, it was possible to separate interpretation and detection errors. ROC analysis was used to compare the performance of the radiologists, and to visualize the relation between their sensitivity and specificity. This experiment is not aimed to be a study on the effect of prompts of the CAD system.

9.3.1 Materials and methods

From the dataset used in the first experiment, 13 malignant abnormalities were selected: 1 architectural distortion, 4 spiculated lesions and 8 vague masses. Five of these lesions were prompted by the system. The set was extended with 107 normal cases. The set resembled screening less than the set in the first experiment because no microcalcifications were present, and easy and obvious malignancies were left out.

A monitor was attached to the alternator, showing the case that was read. With a mouse, the radiologist pointed out all areas they considered as a potential lesion, including regions that are normally discarded immediately. Radiologist marked all regions that they examined in a mammogram, and graded these on a 6-point scale. If an area was graded 1, it was considered very unlikely to be malignant, if it was graded 6 if the radiologist was certain it was a region showing signs of a malignant tumor. After pointing out and grading all the examined regions, the radiologists pressed a button to make the CAD-findings visible. If desired, suspicious regions could be added or deleted.

The radiologists participating in this second experiment were less experienced in mammographic screening than their colleagues in the first experiment. They participated on the last day of a training program for screening mammography, but had extensive clinical mammographic experience before they started the training course.

After the dataset was read, all normal regions that received a high grade and all malignant lesions that received a low grade or that were missed were discussed with the radiologist.

9.3.2 Results

The second experiment confirmed the existence of large variations in the performance of the radiologists. Figure 9.2 shows the ROC curves for the 11 radiologists and a fitted average curve, all radiologists read the same 120 cases. The ROC curve was fit using the ROC analysis software by Metz [7]. The scores are lower than the scores of the 12 radiologists in the first experiment because the radiologists were less experienced and the data set was harder. For each radiologist, a clear relation between his sensitivity and specificity is visible.

In this experiment, 142 times a malignant mass was examined (11 radiologist, 13 abnormalities, 1 accounting mistake). Ten times a tumor was not even assigned grade 1 (a level where on average over 30% of the cases was recalled). These tumors can be considered as misses due to oversight. However, each time the radiologist said he had looked at the region but not even found it worth grade 1. This suggests that the difference between oversight and interpretation is not as clear as one might expect. Whether or not a region was examined is subjective. On the other hand, radiologists won't quickly admit a mistake due to oversight, but rather debate about whether or not a mass is really that suspicious. Because at the least specific operating level approximately 30% of the cases was recalled, we can assume that the regions that were not pointed out were not seriously examined at all, otherwise grade 1 would have been appropriate. Even more interesting than the number of missed tumors is

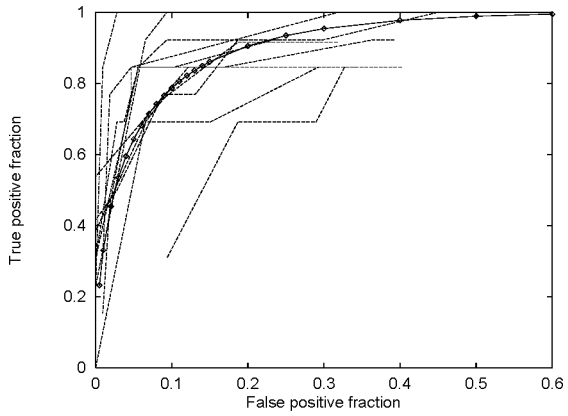

Figure 9.2: ROC curves for the 11 radiologists and a fitted average.

the large number of tumors that was assigned a low grade, lower than many normal regions. These regions obviously were examined, but misinterpreted.

Because in this experiment the findings of the radiologist were recorded directly by the computer, the changes evoked by the prompts could be analyzed easily. Only a few changes were made after the prompts were provided: a few extra cases were sent in yielding a slightly higher sensitivity and slightly lower specificity. Probably one of the reasons for this is the low sensitivity for the masses in the used data set. Two times a tumor was not assigned a measure in the first phase, but did get a value after the prompts are shown. One got the highest measure of suspiciousness, the other a low grade.

9.4 Conclusions

The main conclusion of the first experiment is that the specificity of a CAD-assisted radiologist is not worse than that of a radiologist reading mammograms in a normal way. The number of markers, that exceeded the number of true positive markers by a factor of 50, did not decrease the radiologists' specificity. This confirms the results found in a number of other studies [9, 12].

The inter-observer variability of the radiologists participating in the first experiment, all involved in regular screening, was very high. It has been suggested that radiologists roughly operate on the same ROC curve, i.e. that differences due to training and experience would be small [4]. Ideally, this would be the case, but the results in Table 9.5 show that there is not a clear relation between the sensitivity and specificity of the radiologists. Also after correction for variations in set difficulty there does not seem to be such a relation, indicating that significant differences in skill exist. The second experiment confirms that the ROC curves radiologist work on are quite different. Our results confirm other studies that have shown large variations [10] between radiologists. This large variation in performance may have implications for the prompting device, for example the optimal sensitivity/specificity setting of the software may be vary for different radiologists. Also, it is clear that some radiologist could benefit more from CAD than others.

Most of the cases that were recalled unnecessarily were sent in for a possible malignant mass. Radiologists have problems interpreting and classifying normal tissue. The proportion

of errors due to interpretation seems higher than that of errors due to detection. However, the difference between an error due to interpretation and detection is not as well distinct as sometimes is suggested. The radiologist may have fovially seen the malignant region, but decided it was of no interest on a less-than-conscious level.

One should keep in mind that the average time a radiologist spends on a case in experiments like this is considerable longer than the time that is spend on a case in the screening situation. In screening an experienced radiologist reads over 100 cases per hour, in this study at least 2 hours were used for the 120 cases. Errors due to oversight are therefore more likely to occur in the screening situation and may be hard to find in experiments. Another difference is the target rate, which still is higher in our experiments than in screening situations.

The second experiment shows that when a screening program aims at high specificity levels, this may yield a number of cases showing signs of a cancer that is not detected due to interpretation problems. These cases were not considered suspicious enough to recall, which results in interval carcinomas and late detected cancers.

Bibliography

[1] C A Beam, P M Layde, and D C Sullivan. Variability in the interpretation of screening mammograms by u.s. radiologists. *Arch Intern Med*, 156:209–213, 1996.

[2] H C Burrel, D M Sibbering, A R M Wilson, S E Pinder, A J Evans, L J Yeoman, C W Elston, I O Ellis, R W Blamey, and J F R Robertson. Screening interval breast cancers: mammographic features and prognostic factors. *Radiology*, 199:811–817, 1996.

[3] H P Chan, K Doi, C J Vyborny, R A Schmidt, C E Metz, K L Lam, T Ogura, Y Wu, and H Macmahon. Improvement in radiologist's detection of clustered microcalcifications on mammograms. *Inv Radiol*, 25:1102–1110, 1990.

[4] C J D'Orsi, D J Getty, J A Swets, R M Pickett, S E Seltzer, and B J McNeil. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. *Radiology*, 184(3):519–622, 1992.

[5] J G Elmore, C K Wells, C H Lee, D H Howard, and A R Feinstein. Variability in radiologists' interpretations of mammograms. *N Engl J Med*, 331(22):1493–1499, 1994.

[6] W P Kegelmeyer, J M Pruneda, P D Bourland, A Hillis, M W Riggs, and M L Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191:331–337, 1994.

[7] C E Metz. Evaluation of digital mammography by roc analysis. In K Doi, M L Giger, R M Nishikawa, and R A Schmidt, editors, *Digital Mammography*, pages 61–68. Elsevier, Amsterdam, 1996.

[8] M Moskowitz. Retrospective reviews of breast cancer screening: what do we really learn from them? *Radiology*, 199(3), 1996.

[9] J Roehrig, T Doi, A Hasegawa, B Hunt, J Marshall, H Romsdahl, A Schneider, R Sharbaugh, and W Zang. Clinical results with r2 imagechecker in support of fda pma application. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 395–400. Kluwer, Dordrecht, 1998.

[10] RA Schmidt, GM Newstead, MN Linver, GW Eklund, CE Metz, MN Winkler, and RM Nishikawa. Mammographic screening sensitivity of general radiologists. In

N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 383–388. Kluwer, Dordrecht, 1998.

[11] B Vitak. Invasive interval cancers in the Östergötland mammographic screening programme: Radiological analysis. *European Radiology*, 8:639–646, 1998.

[12] LJ Williams, RJ Prescott, and M Hartswood. Computer-aided cancer detection in the uk breast screening programme. In N Karssemeijer, MAO Thijssen, JHCL Hendriks, and LJTO van Erning, editors, *Digital Mammography*, pages 359–362. Kluwer, Dordrecht, 1998.

Summary and conclusions

This thesis describes the components of an automated detection method for masses and architectural distortions, signs of infiltrating cancer. Masses and architectural distortions can be very subtle and are frequently missed by radiologists. Because the success of treatment of breast cancer depends largely on the stage of the tumor at the time of detection, early detection is very important. Masses have two main image characteristics that can be used for detection: a radiating pattern of spicules and a mass. Sometimes both characteristics are present, but often only spicules or just a faint mass is visible. To achieve high sensitivity on the whole spectrum of possible appearances of masses and distortions, detection of both characteristics is essential. Chapter 2 describes a sensitive method to detect radiating spicule patterns using statistical analysis of line orientations. However, many masses do not show clear spiculation, and must be detected by their mass. Chapter 3 describes how the spicule detection method can be transformed to a mass detection method. Instead of a map of line orientations, a map of gradient orientations is computed. Statistical analysis of this orientation map was used to detect masses. A large set of mammograms taken from the Nijmegen screening program was used to test a detection method based on spicules, a detection method based on masses, and a detection method that detects both spicules and masses. Best results were obtained when both the spiculation and mass features were used. Of all masses, 85% was detected at a specificity level of 1 false positive per image, 55% at 1 false positive per 10 images.

The diameter of masses in mammograms varies from 5 mm to 5 cm, inspiring many research groups to use multi-scale approaches to detect masses. However, the benefit of applying their method in a multi-scale way is almost never compared to a single-scale version of their method. In Chapter 4, the mass detection method of Chapter 3 and two popular pattern recognition techniques to detect bright areas were applied in a single and multi-scale way to examine the possible gain of multi-scale detection. It appeared that the multi-scale versions of the mass detection method had similar performance as a single-scale approach if this scale was chosen appropriately. Of course, when the scale for the single-scale approach was chosen sub-optimally the performance was lower. This study shows that it is not self-evident that a multi-scale mass detection method gives better results than a single-scale version of the method. A multi-scale method is sensitive for masses over a range of sizes, but is also sensitive for false positives of different sizes.

The specificity level that was achieved by the mass detection method described in Chapter 3 is not high enough for successful application in the clinic or in screening. To improve the specificity, a second stage was designed, that classifies each detected region based on regional criteria like contrast, shape, and texture. Based on such features, many normal tissue regions could be discriminated from real masses. To compute these features, a segmentation of the suspicious regions is required. In Chapter 5, a method is described to segment masses

using a discrete dynamic contour model. For each region a size estimate was available of the suspect region, and an appropriate initial starting contour was created that was fitted to the edge of the region. The method proved to be fast and robust, and outperformed a region growing approach. In Chapter 6, the contour model was used to segment regions that were found by the mass detection method of Chapter 3. A number of features were implemented that capture image characteristics that radiologists use to determine whether a suspicious region is a mass or dense normal tissue. Classification using these regional features gave a large reduction in false positives at each desired sensitivity level. On two large datasets a relatively high sensitivity was achieved even at high specificity levels. In Chapter 7, all segmentation methods of Chapter 5 were used to segment and classify the detected regions. The adaptive discrete contour method that was used in Chapter 6 and the preprocessed probabilistic region growing method gave similar results.

The experiments of Chapter 8 showed that a substantial number of the tumors that were missed by radiologists in a screening program despite double reading, were detected by the mass detection method of Chapter 3. Successful detection of missed tumors indicates that a CAD system can be a useful tool for radiologists if the prompts are sufficiently specific. Chapter 9 describes two experiments that were done using a commercially available prompting device. A large experiment showed that the specificity of radiologists does not decrease when they are prompted. This is an important result because some fear that the large number of false positive prompts of a CAD system might increases the recall rate. Results of a second experiment indicated that radiologists have much more difficulty with interpreting suspicious signs than is generally believed. It seems that many screening errors that are thought to be due to oversight, are due to misinterpretation. Both experiments showed large differences in the performance levels of radiologists.

Detection of masses is reaching a level of performance where successful use in screening or clinical practice is possible. Approximately 75% of all masses are detected in at least one view at a specificity level of 0.1 false positives per image. Improvement of the mass and spicule features is still possible, and more sophisticated features can be used to remove false positives. Because the data sets that are used for training are becoming larger, better classifiers can be produced. A considerable improvement can be expected when suspicious regions in one view are correlated to suspicious regions in the other view. Many strong false positives are only present in one of the views, real lesions are most often visible in both. Together with asymmetry features and a method to detect temporal changes in mammograms, another considerable reduction in false positives seems possible.

Samenvatting en conclusies

Dit proefschrift beschrijft de componenten van een automatische detectie methode voor massa's en architecturale verstoringen in mammogrammen. Massa's en verstoringen zijn vaak erg moeilijk te zien en worden daarom regelmatig gemist door radiologen in screening programma's voor borstkanker. Omdat de kans op succesvolle genezing van borstkanker sterk afhangt van de tumorgrootte op het moment van detectie, is vroege opsporing van borstkanker erg belangrijk. Ondersteuning van radiologen bij het opsporen van vroege tekenen van borstkanker door middel van software zou het aantal gemiste tumoren kunnen terugbrengen. Twee belangrijke beeldkenmerken in mammogrammen kunnen gebruikt worden voor automatische detectie van massa's. Het eerste kenmerk is een lijnpatroon rond de tumor, dat wel spiculae genoemd wordt. Het tweede kenmerk is een min of meer rond helder gebied, dat een densiteit of massa genoemd wordt. Soms zijn beide beeldkenmerken aanwezig, vaak is echter slechts 1 van beide zichtbaar. Om het hele spectrum van afwijkingen te kunnen detecteren, moeten beide kenmerken door het algoritme gevonden worden. Hoofdstuk 2 beschrijft een gevoelige methode voor het detecteren van spiculae. De methode maakt een statistische analyse van lijn-richtingen in een gebied, en is daardoor in staat verdachte lijn-patronen te onderscheiden van normaal weefsel. In veel gevallen is het lijnpatroon niet of nauwelijks aanwezig, maar is er wel een densiteit zichtbaar. Hoofdstuk 3 beschrijft hoe de methode om spiculae te detecteren aangepast kan worden om densiteiten te detecteren. In plaats van een analyse van lijn-richtingen wordt een analyse gedaan van gradient-richtingen. Een groot aantal mammogrammen met kwaadaardige afwijkingen die gevonden zijn in het bevolkingsonderzoek naar borstkanker in de regio Nijmegen is gebruikt om de spiculae-en de massa-detectiemethoden te testen. Het beste resultaat werd bereikt wanneer beide beeldkenmerken gebruikt werden in de detectie. Vijfentachtig procent van alle afwijkingen werd gevonden bij een specificiteit van 1 fout-positief per beeld, 55 procent van alle afwijkingen werd nog gevonden bij een specificiteit van 1 fout-positief per 10 beelden.

De diameter van densiteiten in mammogrammen varieert van 5 millimeter tot 5 centimeter. Vanwege deze variatie in grootte hebben veel onderzoeksgroepen gekozen voor een multi-schaal aanpak voor hun massa-detectiemethode. Zelden werd er echter beschreven of deze keuze tot een verbetering leidde ten opzichte van een aanpak met een vaste schaal. In hoofdstuk 4 werden de massa-detectiemethode uit Hoofdstuk 3 en twee populaire patroonherkenningstechnieken om lichte gebieden te detecteren toegepast in een multi-schaal versie en een versie met een vaste schaal. De multi-schaal aanpak bleek even goed te zijn als de aan pak met een vaste schaal, mits deze goed was gekozen. Als deze schaal te klein of te groot gekozen was, was de kwaliteit van de detectiemethode minder. Deze studie laat zien dat het niet vanzelfsprekend is dat een multi-schaal aanpak tot betere resultaten leidt. De multi-schaal aanpak is gevoelig voor densiteiten van verschillende grootte, maar ook voor fout-positieven van verschillende grootte.

De specificiteit van de massa-detectiemethode die beschreven is in hoofdstuk 3 is niet goed genoeg voor gebruik in een klinische omgeving of in een screeningsprogramma naar borstkanker. Om de specificiteit te verbeteren werd de detectiemethode gevolgd door een classificatie programma dat gevonden gebieden classificeerde op basis van regionale kenmerken als grootte, contrast en vorm. Op basis van deze kenmerken konden echte massa's onderscheiden worden van gebieden met normaal weefsel die ook gedetecteerd werden door de massa-detectiemethode. Om dit soort kenmerken te kunnen berekenen, moet het verdachte gebied gesegmenteerd worden. Hoofdstuk 5 beschrijft een methode om gebieden te segmenteren met behulp van een discreet dynamisch contour model. Deze methode bleek robuustere en nauwkeurigere segmentaties van massa's te geven dan een region growing methode. In hoofdstuk 6 is het contour model gebruikt om alle gebieden te segmenteren die gevonden werden door de detectiemethode uit hoofdstuk 3. Een aantal kenmerken werden gedefinieerd die gerelateerd zijn aan beeldkenmerken die radiologen gebruiken om normaal weefsel en echte massa's te onderscheiden. Classificatie van de gevonden gebieden met behulp van deze kenmerken leverde een forse prestatieverbetering op, veel fout-positieve detecties werden verwijderd. Op twee grote databases van mammogrammen werd een hoge sensitiviteit bereikt bij een hoge specificiteit. In hoofdstuk 7 werden verschillende varianten van de region growing methode toegepast op deze sets. De meest geavanceerde versie was even goed als het discreet dynamisch contour model.

De experimenten van hoofdstuk 8 laten zien dat een substantieel gedeelte van de massa's die in het bevolkingsonderzoek werden gemist, gedetecteerd konden worden door de detectiesoftware van hoofdstuk 3. Wanneer gemiste tumoren gedetecteerd worden, kan beslissingsondersteunende software nuttig zijn voor radiologen. Voorwaarde is wel dat de prompts voldoende specifiek zijn, zodat ze serieus worden genomen door de radioloog. Hoofdstuk 9 beschrijft 2 experimenten die gedaan werden met een beschikbaar commercieel systeem dat prompts genereert. Een groot experiment laat zien dat de specificiteit van radiologen niet daalt wanneer ze prompts gebruiken. Dit is een belangrijk resultaat omdat de angst bestond dat door de relatief lage specificiteit van de software het aantal onnodige doorverwijzingen zou stijgen. Een tweede experiment laat zien dat radiologen veel meer problemen hebben met het interpreteren van verdachte gebieden dan algemeen aangenomen werd. Veel fouten in screeningprogramma's worden niet gemaakt omdat radiologen de tumor niet gezien hebben, maar omdat ze de afwijking niet verdacht genoeg vonden om de vrouw door te sturen voor een vervolgonderzoek. Beide experimenten laten zeer grote kwaliteitsverschillen zien tussen de deelnemende radiologen.

Detectie-algoritmen voor massa's bereiken momenteel het kwaliteitsniveau dat succesvolle toepassing in screening en de kliniek mogelijk maakt. Ongeveer 75 procent van alle massa's wordt gedetecteerd bij een specificiteits niveau van 1 fout-positief per 10 beelden. Verbetering van de densiteit en spiculae detectie is mogelijk en betere kenmerken kunnen worden bedacht om fout-positieven te verwijderen. Omdat de databases steeds groter worden, kunnen betere classificatie programma's worden gebouwd. Een grote prestatie verbetering mag worden verwacht wanneer verdachte gebieden in de oblique film gecorreleerd worden met verdachte gebieden in de cranio-caudale film. Veel sterke fout-positieven zijn slechts aanwezig in één van beide projecties, terwijl echte massa's meestal zichtbaar zijn in beide projecties. Samen met asymmetrie kenmerken en een methode om temporele verschillen in mammogrammen op te sporen zal dit aanzienlijke verbetering van de specificiteit tot gevolg hebben.

List of publications

Journal publications

N Karssemeijer, G M te Brake, *Detection of Stellate Distortions in Mammograms*, IEEE Transactions on Medical Imaging, vol. 15, pp 611-619, 1996.

G M te Brake, N Karssemeijer, J H C L Hendriks, *Automated Detection of Breast Carcinomas not Detected in a Screening Program*, Radiology, vol. 207, pp 465-471,1998.

G M te Brake, N Karssemeijer, *Single and Multi-scale Detection of Masses in Digital Mammograms.*, IEEE Transactions on Medical Imaging, vol. 18, nr. 7, July 1999.

G M te Brake, N Karssemeijer, *Segmentation of Suspicious Regions in Mammograms*, Medical Physics. Submitted.

G M te Brake, N Karssemeijer, J H C L Hendriks, *Specificity Improvement by Regional Analysis for Mass Detection Algorithms in Mammograms*, Physics in Medicine and Biology. Submitted.

Publications in conference proceedings

G M te Brake, N Karssemeijer, *Detection Criteria for Evaluation of Computer Aided Diagnosis Systems*, 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam, the Netherlands, 1996.

G M te Brake, N Karssemeijer, *Detection of Stellate Breast Abnormalities*, Digital Mammography, K Doi, M L Giger, R M Nishikawa, R A Schmidt (editors), pp 341-346, Elsevier, Amsterdam, 1996.

N Karssemeijer, G M te Brake, W J H Veldkamp, *Methods in Computer Aided Digital Mammography*, Medical Imaging Technology, Japanese Society of Medical Imaging Technology, Vol. 14, No.6 , 1996.

N Karssemeijer, G M te Brake, *Combining Single View Features and Asymmetry for Detection of Mass Lesions.*, Digital Mammography, N Karssemeijer and M A O Thijssen, J H C L Hendriks, L J T O van Erning (editors), pp 95-102, Kluwer, Dordrecht, 1998.

G M te Brake, N Karssemeijer, *Comparison of Three Mass Detection Methods*, Digital Mammography, N Karssemeijer, M A O Thijssen, J H C L Hendriks, L J T O van Erning (editors), pp 119-126, Kluwer, Dordrecht, 1998.

G M te Brake, M J Stoutjesdijk, N Karssemeijer, *A Discrete Dynamic Contour Model for Mass Segmentation in Digital Mammograms*, Medical Imaging 1999: Image processing, K M Hanson (editor), Vol. 3661, pp 911-919, SPIE, 1999.

Miscellaneous publications

G M te Brake, N Karssemeijer, J H C L Hendriks, *Computer Aided Detection of Masses in Digital Mammograms*, Medical Imaging International. Accepted for publication.

N Karssemeijer, W J H Veldkamp, G M te Brake, J H C L Hendriks, *Beoordeling van Mammogrammen met behulp van Neurale Netwerken*. Nederlands Tijdschrift voor Geneeskunde, 143(45), pp 2232-2236, 6 november 1999.

Dankwoord

Een aantal mensen hebben een grote invloed gehad op het tot stand komen van dit proefschrift. Allereerst wil ik Nico Karssemeijer bedanken voor de uitstekende en prettige begeleiding die ik deze jaren van hem heb gekregen. Ik heb tijdens mijn promotie-onderzoek veel van hem geleerd. Daarnaast ben ik Jan Hendriks dank verschuldigd voor het geduldig uitleggen wat nou wel en wat nou niet verdacht is in mammogrammen. Zijn enthousiasme voor computer-ondersteunde diagnose en zijn hulp bij de experimenten met radiologen hebben veel bijgedragen aan de totstandkoming van de laatste hoofdstukken van dit proefschrift. Stan Gielen wil ik bedanken voor de vriendelijke en behulpzame wijze waarop hij mijn promotor is geweest.

De vele discussies die ik de afgelopen 4 jaar met Wouter Veldkamp heb gehad, zowel in het ziekenhuis als in de Opera, hebben in hoge mate bijgedragen aan dit proefschrift en mijn academische vorming in het algemeen. Mark, Marc en Henk-Jan wil ik bedanken voor de vele gezellige lunches. Henk-Jan, Wouter en Maarten Lamers wil ik bedanken voor het proeflezen van artikelen. Verder wil ik al mijn vrienden en familie bedanken bij wie ik in tijden dat het niet zo soepel liep altijd kon klagen. Bedankt allemaal!

Curriculum Vitae

Guido te Brake is geboren op 9 augustus 1969 te Utrecht. Na het Atheneum te hebben doorlopen op de KSG de Breul in Zeist, ging hij in 1987 Informatica studeren aan de Universiteit Utrecht. In 1991 heeft hij een half jaar gestudeerd aan de University of Wisconsin, Madison. Na in 1993 afgestudeerd te zijn, is hij op 1 januari 1994 de tweejarige postdoctorale opleiding Wiskundige Beheers-en Beleidsmodellen gaan volgen aan de Technische Universiteit Delft. De stage die hij in het kader van deze opleiding heeft gedaan bij de afdeling Radiologie van het Radboud Ziekenhuis Nijmegen, kreeg een vervolg in een promotie-onderzoek. Het resultaat daarvan ligt nu voor u.

R2 Technology, Inc, is proud to sponsor this work as part of its dedication to the improvement of mammography practice by assisting in the detection of early stage breast cancer.

Stereotaxie 2000 D

Digital is Mammographie in a Reality !